Neuron

Article

Metastable attractors explain the variable timing of stable behavioral action sequences

Highlights

- Behavioral sequences in freely moving rats revealed large variability in action timing
- Actions were preceded by onset of specific neural patterns in secondary motor cortex
- Metastable attractors in a network model can explain the origin of timing variability
- Transitions between attractors are driven by low-dimensional correlated variability

Authors

Stefano Recanatesi,
Ulises Pereira-Obilinovic,
Masayoshi Murakami, Zachary Mainen,
Luca Mazzucato

Correspondence

zmainen@neuro.fchampalimaud.org (Z.M.),
lmazzuca@uoregon.edu (L.M.)

In brief

Self-initiated actions in freely moving rats can be predicted by specific ensemble activity patterns in the secondary motor cortex (M2). Variability in action timing can be explained by metastable attractors in a network model of M2. Transitions between attractors are generated by low-dimensional correlated variability, empirically verified in M2.

Recanatesi et al., 2022, Neuron *110*, 139–153
January 5, 2022 © 2021 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.neuron.2021.10.011

CellPress

Neuron

CellPress
OPEN ACCESS

Article

Metastable attractors explain the variable timing of stable behavioral action sequences

Stefano Recanatesi,[1,2,7] Ulises Pereira-Obilinovic,[3,7] Masayoshi Murakami,[4,5] Zachary Mainen,[5,*] and Luca Mazzucato[2,6,8,*]

[1]University of Washington, Center for Computational Neuroscience and Swartz Center, Seattle, WA, USA
[2]Institute of Neuroscience, University of Oregon, Eugene, OR, USA
[3]Center for Neural Science, New York University, New York, NY, USA
[4]Department of Neurophysiology, University of Yamanashi, Yamanashi, Japan
[5]Champalimaud Centre for the Unknown, Lisbon, Portugal
[6]Departments of Biology and Mathematics, University of Oregon, Eugene, OR, USA
[7]These authors contributed equally
[8]Lead contact
*Correspondence: zmainen@neuro.fchampalimaud.org (Z.M.), lmazzuca@uoregon.edu (L.M.)

https://doi.org/10.1016/j.neuron.2021.10.011

SUMMARY

The timing of self-initiated actions shows large variability even when they are executed in stable, well-learned sequences. Could this mix of reliability and stochasticity arise within the same neural circuit? We trained rats to perform a stereotyped sequence of self-initiated actions and recorded neural ensemble activity in secondary motor cortex (M2), which is known to reflect trial-by-trial action-timing fluctuations. Using hidden Markov models, we established a dictionary between activity patterns and actions. We then showed that metastable attractors, representing activity patterns with a reliable sequential structure and large transition timing variability, could be produced by reciprocally coupling a high-dimensional recurrent network and a low-dimensional feedforward one. Transitions between attractors relied on correlated variability in this mesoscale feedback loop, predicting a specific structure of low-dimensional correlations that were empirically verified in M2 recordings. Our results suggest a novel mesoscale network motif based on correlated variability supporting naturalistic animal behavior.

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) (Berman et al., 2016; Wiltschko et al., 2015; Markowitz et al., 2018). Large trial-to-trial variability has been observed in action timing, where transitions between consecutive actions are well described by a Poisson process (Killeen and Fetterman, 1988). Recent studies in *C. elegans* (Linderman et al., 2019), *Drosophila* (Berman et al., 2016), and rodents (Wiltschko et al., 2015; Markowitz et al., 2018) 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 in which the set of actions is rewarded when executed in a fixed order, yet retaining variability in action timing (Murakami et al., 2014, 2017), a hallmark of self-initiated behavior (Killeen and Fetterman, 1988).

Previous studies in rodents have identified the secondary motor cortex (M2) as part of a distributed network involved in motor planning, working memory (Li et al., 2016), and self-initiated tasks (Murakami et al., 2014, 2017). 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 (Inagaki et al., 2019). Are attractor dynamics in M2 confined to delay period activity? Here, we investigate the hypothesis that attractor dynamics can capture the activity of M2 neural circuits in a more naturalistic behavioral setting in which a freely moving animal performs sequences of self-initiated behavior. In particular, we sought to uncover a correspondence between M2 neural activity patterns 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 without the use of time-warping methods, hampering the applicability of traditional trial-averaged measures of neural activity. A principled framework to tackle this issue is to model single-trial neural population dynamics using hidden Markov models (HMMs) (Rabiner, 1989). 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* (Linderman et al., 2019) to rodents (Jones et al., 2007; Mazzucato et al., 2015; Maboudi et al., 2018; La Camera et al., 2019), primates (Gat and Tishby, 1993; Abeles et al., 1995; Ponce-Alvarez et al., 2012; Engel et al., 2016), and humans (Baldassano et al., 2017; Taghia et al., 2018). HMMs 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 HMMs, can be interpreted as metastable attractors, arising from recurrent dynamics in local cortical circuits (Miller and Katz, 2010; Mazzucato et al., 2015). Metastable attractors are produced in biologically plausible network models (Deco and Hugues, 2012; Litwin-Kumar and Doiron, 2012) and have been used to elucidate features of sensory processing (Miller and Katz, 2010; Mazzucato et al., 2015), working memory (Amit and Brunel, 1997), and expectation (Mazzucato et al., 2019) and to explain state-dependent modulations of neural variability (Deco and Hugues, 2012; Litwin-Kumar and Doiron, 2012; Mazzucato et al., 2016). However, although previous models are capable of generating sequential activity (Sompolinsky and Kanter, 1986; Kleinfeld, 1986; Miller and Katz, 2010; Pereira and Brunel, 2020), they are hindered by a fundamental tradeoff between sequence reproducibility and trial-to-trial temporal variability. Namely, they can endogenously generate either reliable sequences without temporal variability (Sompolinsky and Kanter, 1986; Kleinfeld, 1986; Pereira and Brunel, 2020) or, instead, sequences with large temporal variability but unreliable order (Litwin-Kumar and Doiron, 2012; Mazzucato et al., 2015; Treves, 2005). 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 (Murakami et al., 2014, 2017) 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 (Murakami et al., 2014, 2017). 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 such as M2 and subcortical areas such as thalamus (Guo et al., 2018, 2017) and basal ganglia nuclei (Hélie et al., 2015; Desmurget and Turner, 2010; Nakajima et al., 2019) 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 a specific structure of noise correlations (to be low dimensional and aligned between consecutive states in the sequence of neural activations), which we confirmed in the empirical data. Although previous work showed that low-dimensional (differential) correlations in sensory cortex may be detrimental for accurately encoding external stimuli (Moreno-Bote et al., 2014), our results demonstrate that, surprisingly, they are essential for a motor circuit to produce stable yet temporally variable self-initiated action sequences.

RESULTS

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 snouts 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: (1) waiting for a second tone, delivered at random times, then moving to the reward port to collect a large water amount (henceforth referred to as "patient" trials), and (2) terminating the trial at any moment before the second tone, then moving 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 (Figure 1A). The intervals between consecutive actions show large trial-to-trial variability with right-skewed distributions (Figure 1B; Figure S1A), suggestive of a potential stochastic mechanism underlying their action timing (Killeen and Fetterman, 1988).

To uncover the neural correlates of self-initiated actions, we recorded ensemble spike trains from the M2 (from N = 6–20 neurons per session, 9.9 ± 3.6 on average across 33 recorded sessions) of rats engaged in the waiting task (Murakami et al., 2014, 2017). We found that single-trial ensemble neural activity in M2 consistently unfolded through reliable sequences of hidden or latent neural patterns, inferred using a HMM (Figure 1C; Figure 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 (i.e., transitions solely depend on the current pattern), and neurons discharge as Poisson processes with pattern-dependent firing rates. The number of patterns in each session was selected using an unsupervised cross-validation procedure

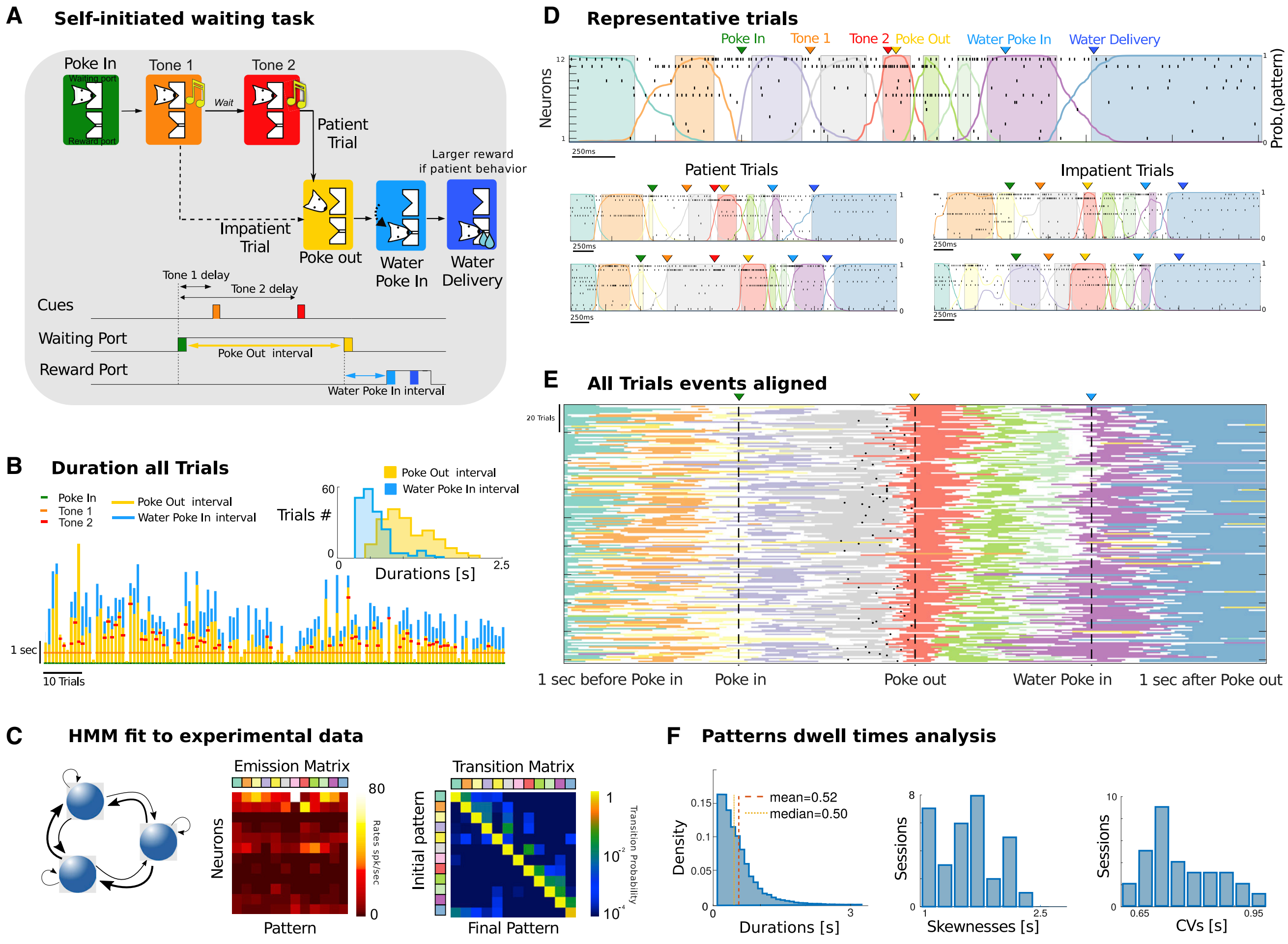

Figure 1. Waiting task and M2 pattern sequences

(A) Schematic of task events. A rat self-initiated the waiting task by poking into a wait port (poke in [PI]), where tone 1 was played (after 400 ms), and after a variable delay, a different tone (tone 2) was played. The animal could decide to poke out [PO] of the wait port at any time (after tone 2 in patient trials, between tones 1 and 2 in impatient trials) and move to the reward port (water poke in [WPI]) 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. Tick marks represent event times (see legend). Vertical bars indicate waiting times for poke out and water poke in (yellow and cyan, respectively). When the red tick (second tone) is not present, that marks an impatient trial. 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, and 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, while impatient trials are displayed but tone 2 is not reported.

(F) Left: histogram of pattern dwell times for all patterns across all 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 (33 sessions).

(10.2 ± 0.6 across 33 sessions, which ranged from 6 to 22 patterns; Figure S2A; STAR Methods) and did not depend on ensemble size (Figure S2B). The identity and order of inferred patterns were remarkably consistent within each session, even across patient and impatient trials. Figure 1D shows five example raster plots in which the sequence of states unfolds through the trial, while Figure 1E shows the sequence of states in all the trials of the same session (Figure S3). The average pattern dwell time was 0.52 ± 0.13 s (mean ± SD across all sessions; dwell time was defined as intervals in which the HMM posterior probability

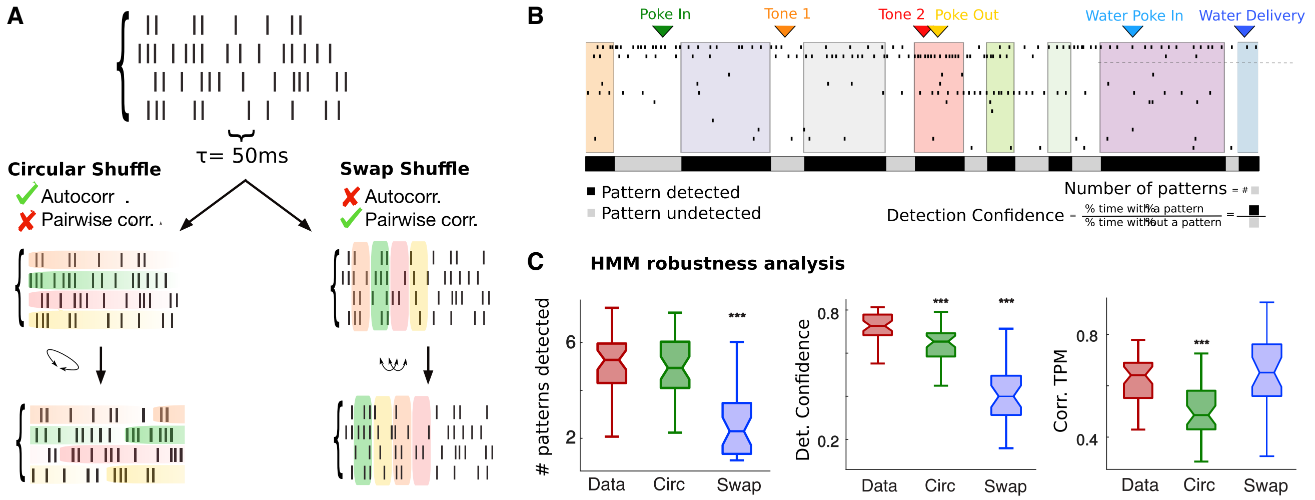

Figure 2. Robustness of pattern inference

(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) Representative trial showing detection confidence measure (same color-coded notation as in Figure 1; black and gray bars, fraction of trial duration during which patterns were detected with probability larger or smaller than 80%, respectively.

(C) HMM robustness analyses. Left: average number of patterns in each trial for empirical and surrogate datasets. Center: pattern detection confidence, estimated as fraction of time across all trials where 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 STAR Methods).

In (A)–(C), signed-rank tests between empirical and shuffled datasets: *p < 0.05, **p < 0.01, and ***p < 0.001.

was above 80%; STAR Methods; Figure 1F; median and interquartile dwell time was 0.42 [0.36–0.49]), in agreement with previous findings in other cortical areas (Mazzucato et al., 2015). Transition intervals between consecutive states lasted 0.14 ± 0.06 s, thus significantly shorter than state durations. 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 (Figure 1F; Figure S2C; coefficient of variability [CV] = 0.76 ± 0.10 and skewness 1.60 ± 0.46). This temporal heterogeneity suggests that a stochastic mechanism may contribute to driving transitions between consecutive patterns within a sequence.

Robustness of pattern inference

We performed a series of control analyses aimed at testing the robustness of our pattern sequence model. We first examined how much single-cell autocorrelation and pairwise correlations contributed to the pattern sequence detection. To do so, first we performed a cross-validation analysis comparing the data with two surrogate datasets (Figure 2A) (Maboudi et al., 2018). 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 with an HMM trained on surrogate datasets (Figures S2F–S2H; empirical versus circular shuffled: $p = 6.5 \times 10^{-7}$; versus swap shuffled: $p = 5.4 \times 10^{-7}$, signed-rank test). When we destroyed autocorrelations, the model entirely failed to detect pattern transitions, leaving only one pattern (Figures 2B and 2C; $p = 5.4 \times 10^{-7}$). 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 (Figure 2C; p = 0.19). However, pattern detection was significantly less confident than in empirical data (Figures 2B and 2C; $p = 3.2 \times 10^{-6}$); moreover, inferred pattern sequences were significantly sparser and more similar across trials in the data compared with the surrogate datasets (Figure 2C; $p = 2.7 \times 10^{-6}$; Figures S2D and S4). We concluded that single-cell autocorrelations, but not pairwise correlations, played an important role in extracting pattern sequences. Moreover, pattern sequences were not driven by the most active neuron in the ensemble, but they are a robust collective property of the whole ensemble dynamics (comparison of HMM fits with the empirical data with surrogate datasets obtained by removing neurons with the highest activity revealed no significant differences; Figures S4B–S4F).

In additional control analyses, we found that the observed neural pattern sequences reflected an underlying discrete process and were not an artifact of the HMM inference. A Gaussian process factor analysis (GPFA; Figure S5; STAR Methods) (Engel et al., 2016; Byron et al., 2008; Churchland and Abbott, 2012) revealed abrupt transitions in the time course of GPFA latent factors separating long and approximately constant epochs with long-tailed dwell time distributions (Figure S5), closely matching

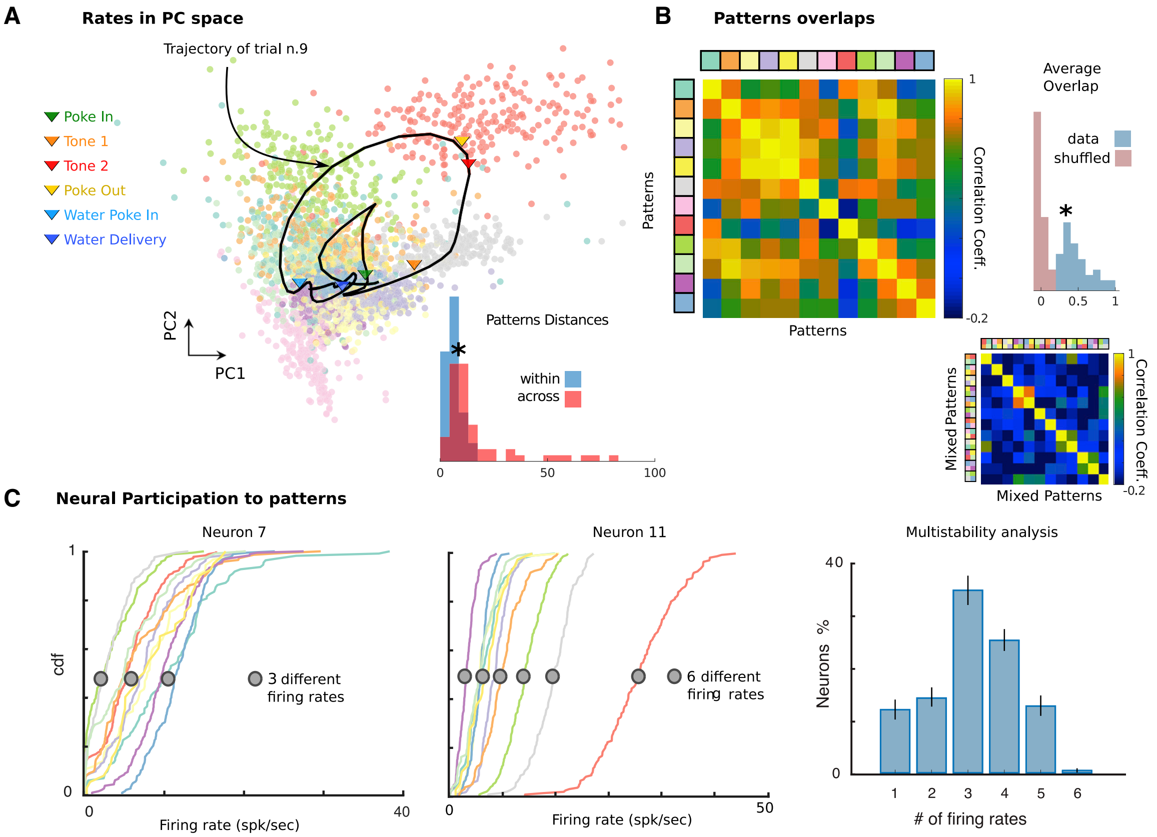

Figure 3. Dense and distributed population code in M2

(A) Neural patterns cluster in principal-component space (all trials from the representative session in Figure 1; color-coded dots represent patterns in single trials; one representative trial smoothed trajectory obtained by averaging neural activity in a sliding window of 600 ms; arrows show events onsets along trajectory). Inset: distribution of within- and across-cluster distances between patterns ($p < 2.0 \times 10^{-7}$, rank-sum test). Colors of different patterns are consistent with previous and following figures where the same example session is analyzed.

(B) Pearson correlation matrix between patterns reveals significantly larger overlaps in the empirical data (top left: representative session) compared with those found when drawing random patterns from the empirical firing rate distribution (bottom right). In this case, the average firing rate of individual neurons in each pattern and trial was randomly drawn from the firing rate distribution of all neurons across all patterns in the same trial. 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 in A and Figure 1D) for two neurons from the representative ensemble, revealing three and six significantly different firing rates across patterns, respectively (see STAR Methods and Figure S2I). Right: number of different firing rates per neuron revealed multistable dynamics where 87% ± 2% of neurons had activities modulated by patterns. Error bars represent the standard error.

the discrete HMM state sequences, and a strong bimodality of GPFA latent factors, consistent with underlying discrete states (Engel et al., 2016).

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 (Figure 3A; $p < 10^{-20}$, Wilcoxon rank-sum test). 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 (Figure 3B; $p = 3 \times 10^{-18}$, t test). We found that the vast majority of neurons (88% ± 2%) had firing rates significantly modulated across patterns (Figure 3C). Although 12% ± 2% of neurons were not modulated and 14% ± 2% had two different firing rates across patterns (bistable neurons), we found that 74% ± 2% of neurons were "multistable," namely, they attained three or more firing rates across all patterns, in agreement with previous findings (Mazzucato et al., 2015). In particular, neurons attained on average 3.22 ± 1.17 different firing rates across patterns underscoring a distributed code of neural patterns across the neural population (Figure 3C; Figure S2I). Such multistability was more pronounced in the empirical data compared with the circular- and swap-shuffled datasets (Figure S4A; Kolmogorov-Smirnov test: empirical versus circular shuffle, $p = 5.9 \times 10^{-6}$; empirical versus circular shuffle, $p = 1.4 \times 10^{-33}$), suggesting that multistability is a property of ensemble dynamics beyond single-neuron autocorrelations and pairwise correlations. Furthermore, we found no linear

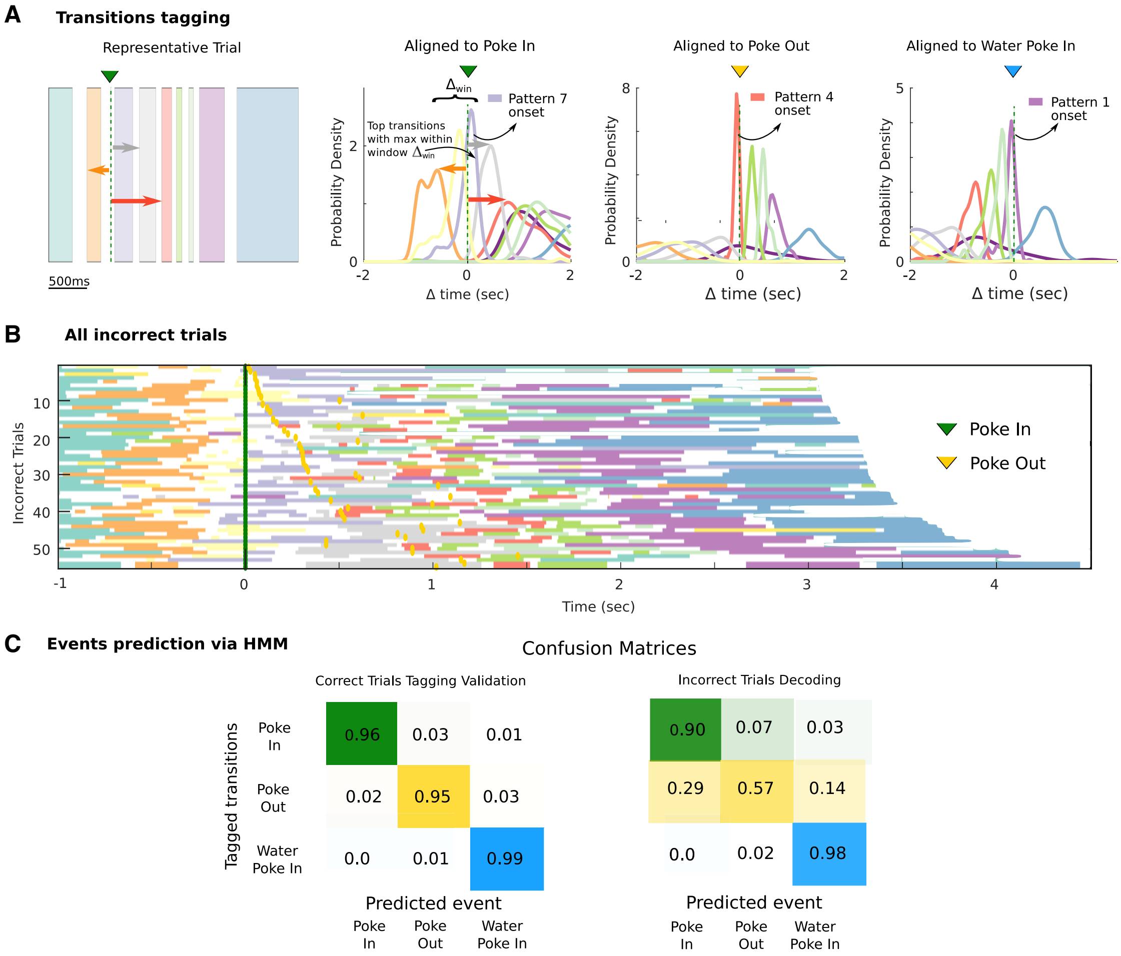

Figure 4. Predicting self-initiated actions from neural pattern onsets

(A) Schematic of pattern/action dictionary. Left: for each action in a correct trial (left: representative trial from Figure 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 Figure 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 (cf. Figure 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 A), action onsets are predicted on test trials (confusion matrix: cross-validated prediction accuracy averaged across 33 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 on the basis of the cross-validated dictionary established in correct trials.

dependence between a single-cell multistability property and its average firing rate (Pearson correlation, $R^2 = 0.11$, p = 0.20), thus suggesting that multistability is unrelated to a cell's average firing rate. 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 several patterns.

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 (Figure 1E; Figure 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 (Figure 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 (−99 ms [−293 to 28 ms], median

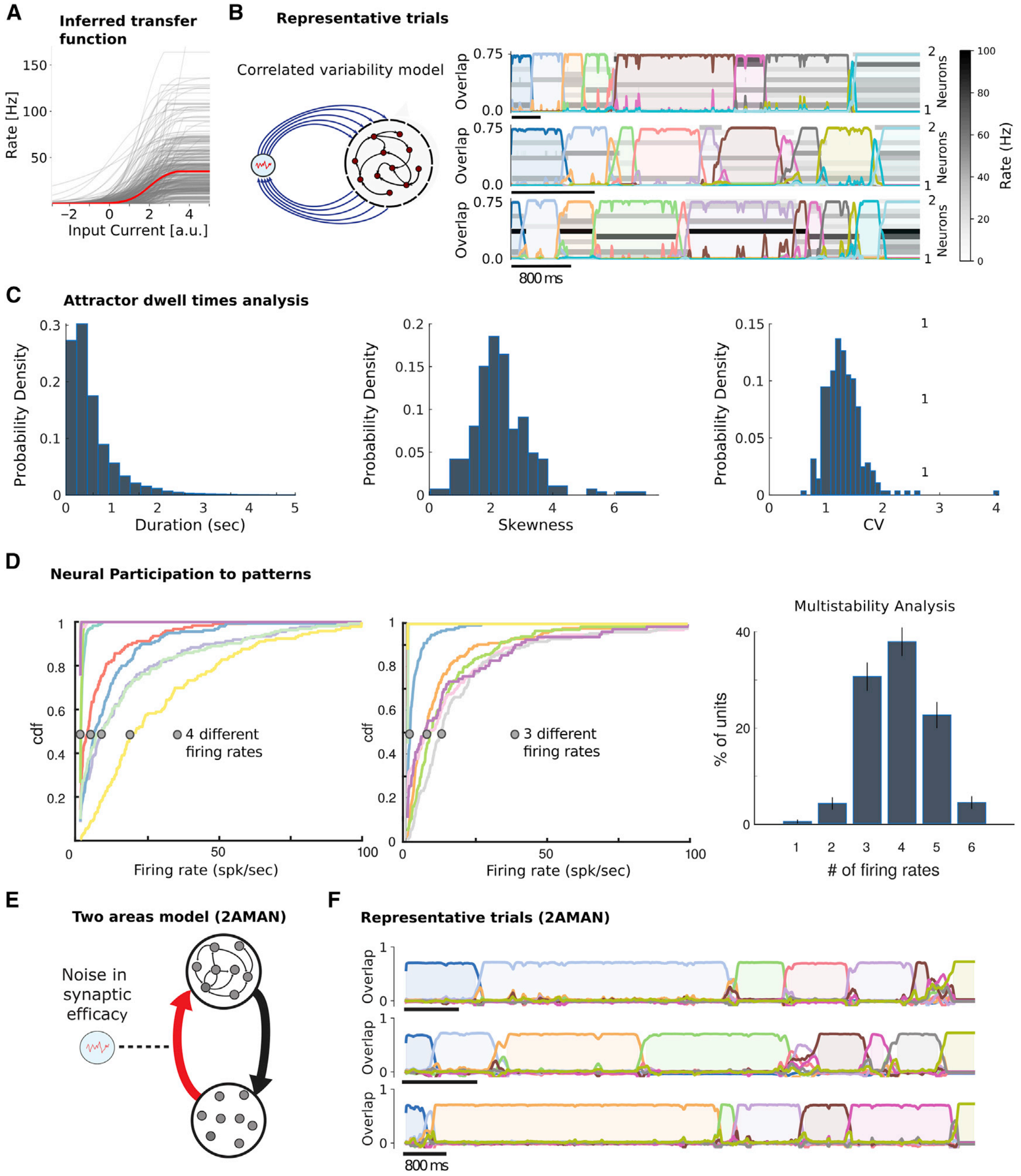

Figure 5. Attractor model of pattern sequences

(A) Distribution of empirical single-cell current-to-rate transfer functions ϕ_i inferred from the data (328 neurons from 33 sessions), used as transfer functions in the recurrent network model (see STAR Methods).

(B) The correlated variability model (Correlated variability originates in a mesoscale feedback loop) 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 Equation 12] 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 patterns 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 (33 simulated networks). The same plots, generated by means of states individuated in the model via a HMM fit on the model simulated neural traces, are shown in Figure S9A.

(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 two and three significantly different firing rates across attractors, respectively (see STAR Methods). Inset: number of different firing rates per neuron revealed multistable dynamics where 99% ± 1% of neurons had activities modulated by patterns.

(legend continued on next page)

[interquartile interval] between pattern onset and tagged action). In correct trials, defined as those in which a visit to the waiting port was followed directly by a movement to the reward port (both patient and impatient types, 67% ± 16% fraction of trials per session; Figure 1A), the cross-validated accuracy of predicting actions from neural patterns was very high (Figure 4C). Trials with other action sequences, in which the animal behaved erratically, were deemed as incorrect trials (Figure 4C; Figure S6).

To assess the significance of the action/pattern dictionary, we aimed at testing whether pattern onsets could correctly predict actions performed during incorrect trials (32.5% ± 15.8%; Figure 4B; Figure S6), on the basis of the dictionary learned solely from correct trials. Pattern sequences in correct trials were more correlated (0.520 ± 0.108, mean ± SD Pearson correlation across sessions) than in incorrect trials (0.398 ± 0.076; $p < 10^{-5}$, t test). The correlation between sequences in correct trials and those in incorrect trials (0.395 ± 0.096) was similar to the correlation found in incorrect trials (p = 0.88, t test). This is consistent with the fact that correct (Figure 1E) and incorrect trials (Figure 4B) both begin with a poke in/poke out and then start diverging. Nevertheless, 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 (Figures 4B and 4C).

Perhaps surprisingly, we found that single neurons in M2 were not responsive to sensory stimuli (auditory tones 1 and 2: 6 and 4 responsive neurons across 328 neurons). Moreover, it was not possible to discriminate patient versus impatient conditions from modulations of population firing rates (from a decoding analysis; Figure S6F) nor from the distribution of pattern dwell times (p > 0.05 in 95% of the sessions, Kolmogorov-Smirnov test), reflecting the consistency of action timing distributions between the two conditions (p > 0.05 in 95% of the sessions). These results are consistent with the hypothesis that M2 neural activity reliably encodes for the animal's actions, regardless of whether these actions are performed in a patient or impatient trial. Indeed, both conditions involved the same action sequence, encoded in a reliable neural pattern sequence occurring in both condition. These results suggest that M2 activity mostly reflected stochasticity in action timing from trial to trial, regardless of whether a trial was classified as patient or impatient. 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.

Correlated variability generates sequences of metastable attractors

What is a possible circuit mechanism underlying the observed pattern sequences? We aimed to capture three main features of the empirical data: (1) long-lived neural patterns (0.5 s on average; Figure 1F), suggesting that they originate from attractor dynamics; (2) right-skewed pattern dwell time distributions (Figure 1F), suggesting that transitions may be noise driven (see, e.g., Gardiner, 1985); and (3) highly reliable sequences across trials (Figure 1E; Figure S3). We thus sought a mechanistic model generating reliable sequences of long-lived attractors with noise-driven transitions between attractors.

The crucial ingredient driving transitions between patterns in the model entails constraining population activity fluctuations along a low-dimensional manifold within a high-dimensional activity space. We achieved this by embedding a low-rank term in the synaptic couplings.

We modeled population activity in M2 as arising from a recurrent network of rate units governed by the following dynamics:

$$\tau \dot{u}_i(t) = -u_i(t) + \sum_{j=1}^{N} J^S_{ij}\phi_j(u_j(t)) + \zeta(t)\sum_{j=1}^{N} J^F_{ij}\phi(u_j(t)),$$

(Equation 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 (fit to the empirical data in M2; Figure 5A; Figure S7). We hypothesized that patterns originated from p discrete attractors η^μ, for $\mu = 1, \ldots, p$, stored in the symmetric synaptic couplings $J^S_{ij} \propto \sum_{\mu=1}^{p} f[\eta_i^\mu] g[\eta_j^\mu]$ (f and g are threshold functions; STAR Methods; Pereira and Brunel, 2018), consistent with experimental evidence supporting discrete attractor dynamics in M2 (Schmitt et al., 2017; Guo et al., 2017; Inagaki et al., 2019). 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 (Figure S7C). Transitions between attractors, giving rise to sequences, originate from the asymmetric term $J^F_{ij} \propto \sum_{\mu=1}^{p} f[\eta_i^{\mu+1}] g[\eta_j^\mu]$ in Equation 1, henceforth referred to as the correlated variability term. This term generates stochastic dynamics via the noise $\zeta(t)$. 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 that are correlated across neurons. Within a large range of parameters (Figure S7D), the network model met all our objectives: (1) long-lived attractors matching the empirical data (average dwell time in Figure 5B fit to the representative session in Figure 1) emerging from the network's collective dynamics; (2) right-skewed dwell time distributions (Figure 5C); and (3) highly reliable attractor sequences (in ~4% of trials, the model generates the wrong sequence of patterns, reminiscent of the incorrect trials in the empirical data). As attractors would be stable in the absence of noise $\zeta(t)$ (Figure S7C), transitions between attractors were entirely noise driven in this model.

(E) Two-area model schematic. Fluctuations in the synaptic efficacy depend only on the pre-synaptic terminals at area Y (see Equation 17) and therefore on the fluctuations on the synaptic efficacy of the Y → M2 synapses.

(F) Three example trials of the two-area model dynamics. As our analytical calculations predict, it produces meta-stable attractor dynamics that quantitatively match our phenomenological model and the data (dwell times distribution not shown). Parameters are the same as in Table 1. The additional parameters take values $N_Y = 1000$, $A_{Y\leftarrow M2} = 0.12$, and $A_{M2\leftarrow Y} = 1$. Area Y's input-output transfer function is the rectifier linear function $\phi(x) = [x-1]_+$.

Furthermore, we found that single-neuron firing rate distributions were heterogeneous (Figure 5D), similar to the empirical ones (Figure 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 (Figure 5C; see also Mazzucato et al., 2015). Compared with traditional attractor models with sparse activations (Tsodyks and Feigel'man, 1988), multistability was accompanied with a more dense code (Figure S8). 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.

Correlated variability originates in a mesoscale feedback loop

The crucial ingredient driving transitions between patterns in the model (see Equation 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 feed-forward circuit (provisionally denoted as Y):

$$\tau \dot{u}_i(t) = -u_i(t) + \sum_{j=1}^{N} J_{ij}^{S} \phi_j(u_j(t)) + \sum_{j=1}^{N_Y} W_{ij}^{M2 \leftarrow Y} r_j \,, \qquad \text{(Equation 2)}$$

$$\tau_Y \dot{r}_i = -r_i + \sum_{j=1}^{N} W_{ij}^{Y \leftarrow M2} \phi_j(u_j) \,.$$

Here, u_i represent the activity of M2 neurons (the same as in Equation 1), and r_i represent activities of neurons in area Y (see model schematic in Figure 5E). The latter area is smaller ($N_Y \ll N$) and faster ($\tau_Y < \tau$) and lacks recurrent couplings, suggesting that it may correspond to a subcortical circuit. The asymmetric term J^F in Equation 1, which generates stochastic transitions between otherwise stable attractors, originates from the reciprocal couplings $W^{Y \leftarrow M2}$ and $W^{M2 \leftarrow Y}$ between M2 and area Y in Equation 2, its temporal dependence arising from fluctuations in the synaptic efficacy of the $Y \rightarrow M2$ synapses (STAR Methods; Figure 5E). The reciprocal connections $W^{Y \leftarrow M2}$ and $W^{M2 \leftarrow Y}$ in this two-area model can be integrated out when dynamics in area Y are faster than in M2 ($\tau_Y < \tau$) (Reinhold et al., 2015; Jaramillo et al., 2019). The two-area mesoscale attractor network (2AMAN) model in Equation 2 is then mathematically equivalent to the effective dynamics in Equation 1, whose recurrent couplings are augmented to include an asymmetric term, J^F, inherited from the reciprocal loop. In the STAR Methods we show how the mean and variance of the noise term $\zeta(t)$ in Equation 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 fluctuations in the synaptic efficacy assuming area Y is small. This variability in the synaptic efficacy may emerge from different but not exclusive cellular mechanisms such as short-term plasticity (Tsodyks et al., 1998) or stochastic synaptic vesicle release (Dobrunz and Stevens, 1997). Network simulations of our two-area model confirm our mathematical results (see Figure 5F).

Correlated variability is necessary to explain temporal variability in attractor networks

Is it possible to generate the observed pattern sequences with alternative mechanisms, in the absence of correlated variability? We varied symmetric (J^S) and asymmetric (J^F) synaptic coupling strength but with no noise ($\zeta(t)$ = const in Equation 1), generating decaying activity, or stable attractors, or sequences of attractors (Figure S7C). However, all these alternative models failed to capture crucial aspects of the data. Namely, dwell time distributions were short, and they showed no trial-to-trial variability, thus being incompatible with the observed patterns (Figure 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 (Figure S8D; STAR 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 (Figure S8E).

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, because of 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 must 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 be driven by noise. We thus concluded that correlated variability, in the context of attractor networks, was necessary to reproduce the right-skewed distribution of pattern dwell time observed in the data.

Low-dimensional variability of M2 pattern sequences

Our recurrent network model (see Equation 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 driving sequences, then two clear predictions should be borne out in the neural population data. First, the correlated variability term in Equation 1 predicts that population activity fluctuations within a given pattern (color-shaded intervals in Figure 5B), henceforth referred to as "noise correlations," lie within a subspace whose dimension is much smaller than that expected by chance (Figure 6A; dimensionality in the model versus shuffled surrogate dataset; $p < 10^{-15}$, rank-sum test). Second, the sequential structure of the correlated variability term in Equation 1 implies that noise correlation directions for attractors that occur in consecutive order within a sequence should be co-aligned. A canonical correlation analysis (CCA) showed that in the correlated variability model the alignment across

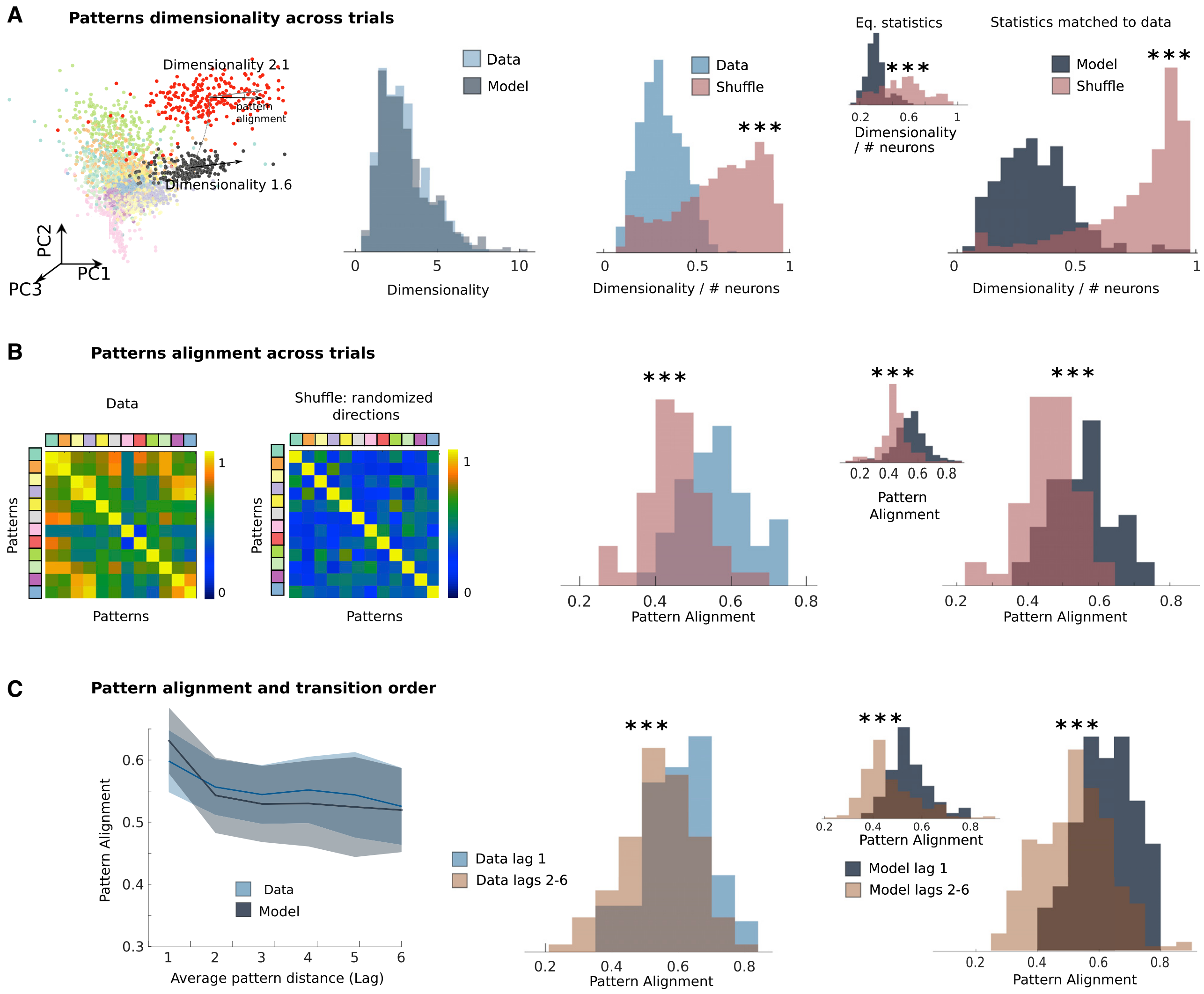

Figure 6. Low-dimensional variability in models and data

(A) Comparison of dimensionality of pattern-conditioned noise correlations in the data (blue) and the model (gray) reveals low-dimensional population activity fluctuations, significantly smaller than expected by chance (red, shuffled datasets). In the shuffle dataset the firing rate of each neuron in each state and trial was randomly sampled from the empirical distributions of firing rates for all states and trials in the same session. From left to right: first panel, representative session as in Figure 1; second panel, summary across 33 sessions from the data and the model; third panel, fractional dimensionality in the data; fourth panel, model's fractional dimensionality estimated by matching ensemble sizes and number trials to data across 33 simulated sessions.

(B) Left panel: 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) both in the data (middle panel) and in the model (right panel).

(C) Left panel: 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 (gray) and in the data (blue). Pattern alignments are significantly larger for patterns at one lag compared with patterns at longer lags both in the data (middle panel) and in the model (right panel). The shaded area corresponds to the standard error.

In all panels, $^*p < 0.05$ and $^{***}p < 0.001$. The same plots in (A)–(C), generated by means of states individuated in the model via a HMM fit on the model's simulated neural traces, are shown in Figures S9B–S9D.

attractor (measured using the top *K* principal components of the noise correlations, where *K* is its dimensionality) was much larger than expected by chance (Figure 6B; alignment in the model versus shuffled surrogate dataset, $p < 10^{-5}$, rank-sum test). More specifically, we found that the strongest alignment occurred between consecutive patterns within a sequence, compared with those occurring further apart (Figure 6B; Figure S9; $p < 10^{-20}$, rank-sum test).

Having established strong statistical features regarding low-dimensional, aligned noise correlations, we tested whether the structure of correlations predicted by the model was 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 (Figure 6A). Applying the same analyses to the data that were run on the model, we found that indeed empirical noise correlations around

Table 1. Network parameters

Model Parameters		
Parameter	Value	Comment
c	0.1	connectivity sparsity
N	10,000	network size
q_f	0.65	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	SD of the synaptic noise
τ_ζ	20 ms	synaptic noise time constant
σ_p	0	SD of the private noise
τ	20 ms	single neuron time constant

each neural pattern had lower dimension than expected by chance and closely matched the dimensionality predicted by the model (Figure 6A). CCA further revealed that noise correlations were highly aligned between patterns, significantly above the alignment expected by chance (Figure 6B; $p = 1.70 \times 10^{-4}$, rank-sum test). Finally, directions of variability were more aligned between consecutive patterns compared with patterns further apart in the sequence ($p < 10^{-14}$, rank-sum test; Figure 6C; Figure S9). Thus, the features of the noise correlations in the neural ensemble data were strongly consistent with the predictions from the correlated variability model.

DISCUSSION

Our results establish a correspondence between self-initiated actions and discrete pattern sequences in 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 mechanistic framework for investigating the neural underpinnings of self-initiated actions and demonstrates a novel link between correlated variability and attractor dynamics.

Evidence for discrete attractor dynamics in cortex

Attractors are characterized by long periods when neural ensembles discharge persistently at approximately constant firing rate (defining a neural pattern) punctuated by relatively abrupt transitions to a different relatively constant pattern. Evidence for attractors was reported in temporal (Fuster and Jervey, 1981; Miyashita, 1988) and frontal areas in primates (Fuster and Alexander, 1971; Funahashi et al., 1989) and rodents (Erlich et al., 2011; Schmitt et al., 2017; Guo et al., 2017; Inagaki et al., 2019), and in rodent sensory cortex (Jones et al., 2007; Ponce-Alvarez et al., 2012; Mazzucato et al., 2015).

Experimental evidence for stimulus-driven sequences of metastable attractors was previously found in primate frontal areas (Gat and Tishby, 1993; Abeles et al., 1995; Seidemann et al., 1996) and rodent sensory areas (Jones et al., 2007). Random sequences were also observed during ongoing periods (Mazzucato et al., 2015, 2016; Engel et al., 2016). 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 (Miller and Katz, 2010; Mazzucato et al., 2015, 2019)(Wyrick and Mazzucato, 2021). A novel feature of our results is that the reliable sequence of metastable attractors is not driven by external stimuli but rather is internally generated.

Neural circuits underlying pattern sequences

The main features of M2 ensemble activity explained by our network models were the reliable identity and order of long-lived neural patterns occurring in a sequence, and the large trial-to-trial variability of pattern dwell times. Both features can be robustly attained when transitions between attractors arise from correlated variability. For an extended comparison with other mechanistic models of attractor sequences, see Table S1.

How does our 2AMAN model architecture map onto specific neural circuits? Previous work showed, using inactivation experiments, that the stochastic component in action timing variability originated in M2 (Murakami et al., 2017). Our 2AMAN model relies on a small and fast network lacking recurrent couplings, representing a subcortical circuit connected to M2, such as the areas that constitute its basal ganglia or thalamic nuclei, as suggested by recent perturbation experiments (Guo et al., 2017; Schmitt et al., 2017). Although a larger mesoscale network may underlie sequence generation, including cortex, thalamus, and basal ganglia (Kawai et al., 2015; Hélie et al., 2015; Desmurget and Turner, 2010; Nakajima et al., 2019; Markowitz et al., 2018; Murray and Escola, 2017; Nakajima et al., 2019; Kao et al., 2005) or a distributed mesoscale network (Svoboda and Li, 2018).

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 (Erlich et al., 2011; Li et al., 2015; Chen et al., 2017; Sul et al., 2011; Inagaki et al., 2019; Guo et al., 2014) as well as self-initiated tasks (Murakami et al., 2014, 2017). The pattern sequences we uncovered in M2 were consistent with the features of preparatory activity (Jin and Costa, 2015): a precise dictionary linked specific patterns to actions, pattern onset reliably predicted action onset, and action timing variability strongly correlated with pattern onset variability.

Correlated variability in sensory versus 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 (Darshan et al., 2017). Low-dimensional correlated variability has been widely reported in sensory cortex, where it may carry information about the animal's behavioral state (McGinley et al., 2015; Cohen and Maunsell, 2009; Huang et al., 2019) or movements (Niell and Stryker, 2010; Polack et al., 2013; Stringer et al., 2019; Musall et al., 2019; Salkoff et al., 2020). It has been proposed that low-dimensional correlated variability in sensory cortex may be detrimental to sensory processing, as it may limit a network's information processing capability (Moreno-Bote et al., 2014). Here, we found that low-dimensional correlated fluctuations in a motor area 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 to avoid predation or competition or to explore the temporal aspects of a given 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 (Kao et al., 2005; Goldberg and Fee, 2011; Darshan et al., 2017). Our results thus suggest that low-dimensional correlations are essential for motor generation.

Action timing variability and M2 activity

Here, we provided new evidence suggesting that M2 is involved in generating self-initiated actions and, crucially, show how variability in M2 population dynamics could generate the variability in self-initiated behavior. Our previous studies showed that most of the variance in waiting times of impatient trials is of stochastic origin, although a small fraction is of deterministic origin and could be interpreted as a trial history-dependent bias in waiting times (Murakami et al., 2014, 2017), likely originating in the prefrontal cortex (PFC). Waiting time itself, however, was encoded only in M2 activity, not in PFC, consistent with the parsimonious hypothesis that the stochasticity in waiting time originates in a circuit that includes M2 itself. These results thus provide strong evidence for our interpretation that M2 circuits are directly involved in the decision of which action to plan and when to act, generating the stochasticity in self-initiated action timing. More generally, these results support a combined PFC-M2-subcortical picture leading to the decision to act: PFC provides a deterministic choice bias, which is translated into an actual choice signal by a downstream circuit including M2, injecting stochastic trial-to-trial variability through a subcortical feedback loop.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2021.10.011.

ACKNOWLEDGMENTS

We would like to thank Fanny Cazettes, Nick Steinmetz, Dennis Nestvogel, and Yonathan Loewenstein for discussions. We thank in particular Laureline Logiaco, James M. Murray, and Nicolas Brunel for suggestions on the network model. L.M. was supported by National Institute of Deafness and Other Communication Disorders grant K25-DC013557, by National Institute of Neurological Disorders and Stroke grant R01-NS118461 (BRAIN Initiative), and by The Swartz Foundation award 66438. Z.M. was supported by the European Research Council (671251) and by the Champalimaud Foundation. M.M. and Z.M. were supported by the Fundação para a Ciência e a Tecnologia (FCT-PTDC/MED-NEU/32068/2017). M.M. was supported by the MEXT Grant-in-Aid for Scientific Research (19H05208, 19H05310, 19K06882) and the Takeda Science Foundation. S.R. and U.P.-O. were supported by The Swartz Foundation.

AUTHOR CONTRIBUTIONS

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

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 10, 2021
Revised: August 30, 2021
Accepted: October 5, 2021
Published: October 29, 2021

SUPPORTING CITATIONS

The following references appear in the supplemental information: Abeles (1991); Dehaene et al. (1987); Diesmann et al. (1999); Fiete et al. (2010);

Hahnloser et al. (2002); Jun and Jin (2007); Liu and Buonomano (2009); Mongillo et al. (2003); Nádasdy et al. (1999); Rajan et al. (2016).

REFERENCES

Abbott, L.F., Rajan, K., and Sompolinsky, H. (2011). Interactions between intrinsic and stimulus-evoked activity in recurrent neural networks. In The Dynamic Brain: An Exploration of Neuronal Variability and Its Functional Significance, M. Ding and D. Glanzman, eds. (Oxford, UK: Oxford University Press).

Abeles, M. (1991). Corticonics (New York: Cambridge University Press).

Abeles, M., Bergman, H., Gat, I., Meilijson, I., Seidemann, E., Tishby, N., and Vaadia, E. (1995). Cortical activity flips among quasi-stationary states. Proc. Natl. Acad. Sci. U S A *92*, 8616–8620.

Amit, D.J., and Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cereb. Cortex *7*, 237–252.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J.W., Hasson, U., and Norman, K.A. (2017). Discovering event structure in continuous narrative perception and memory. Neuron *95*, 709–721.e5.

Berman, G.J., Bialek, W., and Shaevitz, J.W. (2016). Predictability and hierarchy in Drosophila behavior. Proc. Natl. Acad. Sci. U S A *113*, 11943–11948.

Byron, M.Y., Cunningham, J.P., Santhanam, G., Ryu, S.I., Shenoy, K.V., and Sahani, M. (2008). Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. Adv. Neural Inform. Process. Syst. *21*, 1881–1888.

Chen, T.-W., Li, N., Daie, K., and Svoboda, K. (2017). A map of anticipatory activity in mouse motor cortex. Neuron *94*, 866–879.e4.

Churchland, M.M., and Abbott, L.F. (2012). Two layers of neural variability. Nat. Neurosci. *15*, 1472–1474.

Cohen, M.R., and Maunsell, J.H. (2009). Attention improves performance primarily by reducing interneuronal correlations. Nat. Neurosci. *12*, 1594–1600.

Darshan, R., Wood, W.E., Peters, S., Leblois, A., and Hansel, D. (2017). A canonical neural mechanism for behavioral variability. Nat. Commun. *8*, 15415.

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

Dehaene, S., Changeux, J.-P., and Nadal, J.-P. (1987). Neural networks that learn temporal sequences by selection. Proc. Natl. Acad. Sci. U S A *84*, 2727–2731.

Desmurget, M., and Turner, R.S. (2010). Motor sequences and the basal ganglia: kinematics, not habits. J. Neurosci. *30*, 7685–7690.

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

Dobrunz, L.E., and Stevens, C.F. (1997). Heterogeneity of release probability, facilitation, and depletion at central synapses. Neuron *18*, 995–1008.

Domany, E., Hemmen, J.v., and Schulten, K. (1995). Models of Neural Networks I (New York: Springer).

Engel, T.A., Steinmetz, N.A., Gieselmann, M.A., Thiele, A., Moore, T., and Boahen, K. (2016). Selective modulation of cortical state during spatial attention. Science *354*, 1140–1144.

Erlich, J.C., Bialek, M., and Brody, C.D. (2011). A cortical substrate for memory-guided orienting in the rat. Neuron *72*, 330–343.

Fiete, I.R., Senn, W., Wang, C.Z., and Hahnloser, R.H. (2010). Spike-time-dependent plasticity and heterosynaptic competition organize networks to produce long scale-free sequences of neural activity. Neuron *65*, 563–576.

Funahashi, S., Bruce, C.J., and Goldman-Rakic, P.S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. J. Neurophysiol. *61*, 331–349.

Fuster, J.M., and Alexander, G.E. (1971). Neuron activity related to short-term memory. Science *173*, 652–654.

Fuster, J.M., and Jervey, J.P. (1981). Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. Science *212*, 952–955.

Gale, D., and Shapley, L.S. (2013). College admissions and the stability of marriage. Am. Math. Mon. *120*, 386–391.

Gardiner, C.W. (1985). Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences. 3rd (Berlin, Germany: Springer).

Gat, I., and Tishby, N. (1993). Statistical modeling of cell assemblies activities in associative cortex of behaving monkeys. Adv. Neural Inform. Process. Syst *5*, 945–952.

Gillett, M., Pereira, U., and Brunel, N. (2020). Characteristics of sequential activity in networks with temporally asymmetric Hebbian learning. Proc. Natl. Acad. Sci. U S A *117*, 29948–29958.

Goldberg, J.H., and Fee, M.S. (2011). Vocal babbling in songbirds requires the basal ganglia-recipient motor thalamus but not the basal ganglia. J. Neurophysiol. *105*, 2729–2739.

Grossberg, S. (1969). On learning and energy-entropy dependence in recurrent and nonrecurrent signed networks. J. Stat. Phys. *1*, 319–350.

Guo, Z.V., Li, N., Huber, D., Ophir, E., Gutnisky, D., Ting, J.T., Feng, G., and Svoboda, K. (2014). Flow of cortical activity underlying a tactile decision in mice. Neuron *81*, 179–194.

Guo, Z.V., Inagaki, H.K., Daie, K., Druckmann, S., Gerfen, C.R., and Svoboda, K. (2017). Maintenance of persistent activity in a frontal thalamocortical loop. Nature *545*, 181–186.

Guo, K., Yamawaki, N., Svoboda, K., and Shepherd, G.M.G. (2018). 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. J. Neurosci. *38*, 8787–8797.

Hahnloser, R.H., Kozhevnikov, A.A., and Fee, M.S. (2002). An ultra-sparse code underlies the generation of neural sequences in a songbird. Nature *419*, 65–70.

Hélie, S., Ell, S.W., and Ashby, F.G. (2015). Learning robust cortico-cortical associations with the basal ganglia: an integrative review. Cortex *64*, 123–135.

Herz, A., Sulzer, B., Kühn, R., and van Hemmen, J.L. (1989). Hebbian learning reconsidered: representation of static and dynamic objects in associative neural nets. Biol. Cybern. *60*, 457–467.

Holmgren, C., Harkany, T., Svennenfors, B., and Zilberter, Y. (2003). Pyramidal cell communication within local networks in layer 2/3 of rat neocortex. J. Physiol. *551*, 139–153.

Huang, C., Ruff, D.A., Pyle, R., Rosenbaum, R., Cohen, M.R., and Doiron, B. (2019). Circuit models of low-dimensional shared variability in cortical networks. Neuron *101*, 337–348.e4.

Inagaki, H.K., Fontolan, L., Romani, S., and Svoboda, K. (2019). Discrete attractor dynamics underlies persistent activity in the frontal cortex. Nature *566*, 212–217.

Jaramillo, J., Mejias, J.F., and Wang, X.-J. (2019). Engagement of pulvino-cortical feedforward and feedback pathways in cognitive computations. Neuron *101*, 321–336.e9.

Jin, X., and Costa, R.M. (2015). Shaping action sequences in basal ganglia circuits. Curr. Opin. Neurobiol. *33*, 188–196.

Jones, L.M., Fontanini, A., Sadacca, B.F., Miller, P., and Katz, D.B. (2007). Natural stimuli evoke dynamic sequences of states in sensory cortical ensembles. Proc. Natl. Acad. Sci. U S A *104*, 18772–18777.

Jun, J.K., and Jin, D.Z. (2007). Development of neural circuitry for precise temporal sequences through spontaneous activity, axon remodeling, and synaptic plasticity. PLoS ONE *2*, e723.

Kao, M.H., Doupe, A.J., and Brainard, M.S. (2005). Contributions of an avian basal ganglia-forebrain circuit to real-time modulation of song. Nature *433*, 638–643.

Kawai, R., Markman, T., Poddar, R., Ko, R., Fantana, A.L., Dhawale, A.K., Kampff, A.R., and Ölveczky, B.P. (2015). Motor cortex is required for learning but not for executing a motor skill. Neuron *86*, 800–812.

Killeen, P.R., and Fetterman, J.G. (1988). A behavioral theory of timing. Psychol. Rev. *95*, 274–295.

Kleinfeld, D. (1986). Sequential state generation by model neural networks. Proc. Natl. Acad. Sci. U S A *83*, 9469–9473.

La Camera, G., Fontanini, A., and Mazzucato, L. (2019). Cortical computations via metastable activity. Current opinion in neurobiology *58*, 37–45. https://doi.org/10.1016/j.conb.2019.06.007.

Lefort, S., Tomm, C., Floyd Sarria, J.C., and Petersen, C.C. (2009). The excitatory neuronal network of the C2 barrel column in mouse primary somatosensory cortex. Neuron *61*, 301–316.

Li, N., Chen, T.-W., Guo, Z.V., Gerfen, C.R., and Svoboda, K. (2015). A motor cortex circuit for motor planning and movement. Nature *519*, 51–56.

Li, N., Daie, K., Svoboda, K., and Druckmann, S. (2016). Robust neuronal dynamics in premotor cortex during motor planning. Nature *532*, 459–464.

Lim, S., McKee, J.L., Woloszyn, L., Amit, Y., Freedman, D.J., Sheinberg, D.L., and Brunel, N. (2015). Inferring learning rules from distributions of firing rates in cortical neurons. Nat. Neurosci. *18*, 1804–1810.

Linderman, S.W., Nichols, A.L., Blei, D.M., Zimmer, M., and Paninski, L. (2019). Hierarchical recurrent state space models reveal discrete and continuous dynamics of neural activity in C. elegans. bioRxiv. https://doi.org/10.1101/621540.

Litwin-Kumar, A., and Doiron, B. (2012). Slow dynamics and high variability in balanced cortical networks with clustered connections. Nat. Neurosci. *15*, 1498–1505.

Liu, J.K., and Buonomano, D.V. (2009). Embedding multiple trajectories in simulated recurrent neural networks in a self-organizing manner. J. Neurosci. *29*, 13172–13181.

Maboudi, K., Ackermann, E., de Jong, L.W., Pfeiffer, B.E., Foster, D., Diba, K., and Kemere, C. (2018). Uncovering temporal structure in hippocampal output patterns. eLife *7*, e34467.

Markowitz, J.E., Gillis, W.F., Beron, C.C., Neufeld, S.Q., Robertson, K., Bhagat, N.D., Peterson, R.E., Peterson, E., Hyun, M., Linderman, S.W., et al. (2018). The striatum organizes 3d behavior via moment-to-moment action selection. Cell *174*, 44–58.e17.

Markram, H., Lübke, J., Frotscher, M., Roth, A., and Sakmann, B. (1997). Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex. J. Physiol. *500*, 409–440.

Mason, A., Nicoll, A., and Stratford, K. (1991). Synaptic transmission between individual pyramidal neurons of the rat visual cortex in vitro. J. Neurosci. *11*, 72–84.

Mazzucato, L., Fontanini, A., and La Camera, G. (2015). Dynamics of multistable states during ongoing and evoked cortical activity. J. Neurosci. *35*, 8214–8231.

Mazzucato, L., Fontanini, A., and La Camera, G. (2016). Stimuli reduce the dimensionality of cortical activity. Front. Syst. Neurosci. *10*, 11.

Mazzucato, L., La Camera, G., and Fontanini, A. (2019). Expectation-induced modulation of metastable activity underlies faster coding of sensory stimuli. Nat. Neurosci. *22*, 787–796.

McGinley, M.J., Vinck, M., Reimer, J., Batista-Brito, R., Zagha, E., Cadwell, C.R., Tolias, A.S., Cardin, J.A., and McCormick, D.A. (2015). Waking state: rapid variations modulate neural and behavioral responses. Neuron *87*, 1143–1161.

Miller, K.D., and Fumarola, F. (2012). Mathematical equivalence of two common forms of firing rate models of neural networks. Neural Comput. *24*, 25–31.

Miller, P., and Katz, D.B. (2010). Stochastic transitions between neural states in taste processing and decision-making. J. Neurosci. *30*, 2559–2570.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature *335*, 817–820.

Mongillo, G., Amit, D.J., and Brunel, N. (2003). Retrospective and prospective persistent activity induced by Hebbian learning in a recurrent cortical network. Eur. J. Neurosci. *18*, 2011–2024.

Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P., and Pouget, A. (2014). Information-limiting correlations. Nat. Neurosci. *17*, 1410–1417.

Murakami, M., Vicente, M.I., Costa, G.M., and Mainen, Z.F. (2014). Neural antecedents of self-initiated actions in secondary motor cortex. Nat. Neurosci. *17*, 1574–1582.

Murakami, M., Shteingart, H., Loewenstein, Y., and Mainen, Z.F. (2017). Distinct sources of deterministic and stochastic components of action timing decisions in rodent frontal cortex. Neuron *94*, 908–919.e7.

Murray, J.M., and Escola, G.S. (2017). Learning multiple variable-speed sequences in striatum via cortical tutoring. eLife *6*, e26084.

Musall, S., Kaufman, M.T., Juavinett, A.L., Gluf, S., and Churchland, A.K. (2019). Single-trial neural dynamics are dominated by richly varied movements. Nature neuroscience *22* (10), 1677–1686.

Nádasdy, Z., Hirase, H., Czurkó, A., Csicsvari, J., and Buzsáki, G. (1999). Replay and time compression of recurring spike sequences in the hippocampus. J. Neurosci. *19*, 9497–9507.

Nakajima, M., Schmitt, L.I., and Halassa, M.M. (2019). Prefrontal cortex regulates sensory filtering through a basal ganglia-to-thalamus pathway. Neuron *103*, 445–458.e10.

Niell, C.M., and Stryker, M.P. (2010). Modulation of visual responses by behavioral state in mouse visual cortex. Neuron *65*, 472–479.

Pereira, U., and Brunel, N. (2018). Attractor dynamics in networks with learning rules inferred from in vivo data. Neuron *99*, 227–238.e4.

Pereira, U., and Brunel, N. (2020). Unsupervised learning of persistent and sequential activity. Front. Comput. Neurosci. *13*, 97.

Polack, P.-O., Friedman, J., and Golshani, P. (2013). Cellular mechanisms of brain state-dependent gain modulation in visual cortex. Nat. Neurosci. *16*, 1331–1339.

Ponce-Alvarez, A., Nácher, V., Luna, R., Riehle, A., and Romo, R. (2012). Dynamics of cortical neuronal ensembles transit from decision making to storage for later report. J. Neurosci. *32*, 11956–11969.

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE *77*, 257–286.

Rajan, K., Harvey, C.D., and Tank, D.W. (2016). Recurrent network models of sequence generation and memory. Neuron *90*, 128–142.

Reinhold, K., Lien, A.D., and Scanziani, M. (2015). Distinct recurrent versus afferent dynamics in cortical visual processing. Nat. Neurosci. *18*, 1789–1797.

Salkoff, D.B., Zagha, E., McCarthy, E., and McCormick, D.A. (2020). Movement and performance predict widespread cortical activity in a visual detection task. Cerebral Cortex *30* (1), 421–437.

Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a "kneedle" in a haystack: detecting knee points in system behavior. In 2011 31st International Conference on Distributed Computing Systems Workshops, pp. 166–171.

Schmitt, L.I., Wimmer, R.D., Nakajima, M., Happ, M., Mofakham, S., and Halassa, M.M. (2017). Thalamic amplification of cortical connectivity sustains attentional control. Nature *545*, 219–223.

Seidemann, E., Meilijson, I., Abeles, M., Bergman, H., and Vaadia, E. (1996). 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*, 752–768.

Sompolinsky, H., and Kanter, I. (1986). Temporal association in asymmetric neural networks. Phys. Rev. Lett. *57*, 2861–2864.

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C.B., Carandini, M., and Harris, K.D. (2019). Spontaneous behaviors drive multidimensional, brainwide population activity. Science *364*. https://doi.org/10.1126/science.aav7893.

Sul, J.H., Jo, S., Lee, D., and Jung, M.W. (2011). Role of rodent secondary motor cortex in value-based action selection. Nat. Neurosci. *14*, 1202–1208.

Svoboda, K., and Li, N. (2018). Neural mechanisms of movement planning: motor cortex and beyond. Curr. Opin. Neurobiol. *49*, 33–41.

Taghia, J., Cai, W., Ryali, S., Kochalka, J., Nicholas, J., Chen, T., and Menon, V. (2018). Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition. Nat. Commun. *9*, 2505.

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

Tomasi, G., and Bro, R. (2006). A comparison of algorithms for fitting the parafac model. Comput. Stat. Data Anal. *50*, 1700–1734.

Treves, A. (2005). Frontal latching networks: a possible neural basis for infinite recursion. Cogn. Neuropsychol. *22*, 276–291.

Tsodyks, M.V., and Feigel'man, M.V. (1988). The enhanced storage capacity in neural networks with low activity level. EPL *6*, 101.

Tsodyks, M.V., and Markram, H. (1997). The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. Proc. Natl. Acad. Sci. U S A *94*, 719–723.

Tsodyks, M., Pawelzik, K., and Markram, H. (1998). Neural networks with dynamic synapses. Neural Comput. *10*, 821–835.

Williams, A.H., Kim, T.H., Wang, F., Vyas, S., Ryu, S.I., Shenoy, K.V., Schnitzer, M., Kolda, T.G., and Ganguli, S. (2018). Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis. Neuron *98*, 1099–1115.e8.

Wiltschko, A.B., Johnson, M.J., Iurilli, G., Peterson, R.E., Katon, J.M., Pashkovski, S.L., Abraira, V.E., Adams, R.P., and Datta, S.R. (2015). Mapping sub-second structure in mouse behavior. Neuron *88*, 1121–1135.

Wyrick, D., and Mazzucato, L. (2021). State-dependent regulation of cortical processing speed via gain modulation. Journal of Neuroscience *41*, 3988–4005.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental models: Organisms/strains		
Long-Evans hooded rats.	Harlan	Le/CpbHsd
Software and algorithms		
MATLAB	MathWorks	MATLAB R2019a
Python version 3.7	Python	https://www.python.org/
Custom code for network simulations	This paper	https://github.com/ulisespereira/sequences-attractors-M2
MClust-3.5	A. David Radish, University of Minnesota	http://redishlab.neuroscience.umn.edu/MClust/MClust.html
Other		
Behavior control system: BControl	Carlos D. Brody, Princeton University	https://brodylabwiki.princeton.edu/bcontrol/index.php/Main_Page
Tetrode wire	H.P.Reid	Polyimide-insulated nichrome 0.0005'' diameter
NSpike data acquisition system	L.M. Frank, University of California, San Francisco, and J. MacArthur, Harvard University Electronic Instrument Design Lab	http://nspike.sourceforge.net/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the corresponding authors, Zachary Mainen (zmainen@neuro.fchampalimaud.org) and Luca Mazzucato (lmazzuca@uoregon.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The code for simulating the network model is available at the following GitHub repository https://github.com/ulisespereira/sequences-attractors-M2. Data or data analysis scripts are available upon reasonable request from the corresponding authors.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All procedures involving animals were either carried out in accordance with US National Institutes of Health standards and approved by Cold Spring Harbor Laboratory Institutional Animal Care and Use Committee or in accordance with European Union Directive 86/609/EEC and approved by Direção-Geral de Veterinária. Experiments were performed on 37 male adult Long-Evans hooded rats. Rats had free access to food, but water was restricted to the behavioral session and 20˜30 additional min per day. Animals were involved in previous procedures.

Behavioral task

Rats were trained on the self-initiated waiting task (Figure 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 ("Poke In"). If the rat stayed in the Wait port for T1 delay (0.4 s), 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 0.5 s 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 0.5 s delay (impatient trial). The rat had to visit the Reward port within 2 s 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 0.7 s 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 Poke In 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 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.” 5 animals were trained on only the delay variant of the task. 3 animals where trained on both variants (delay and lever). We restricted the statistics of recordings to sessions where 6 or more neurons were simultaneously recorded. Upon applying such criteria a total of 33 sessions including 33 delay blocks and 21 lever blocks were analyzed across 8 animals. Each block last for 70-100 trials. Transitions between the blocks were not signaled. 33 sessions (7 rats) were recorded, see Murakami et al., 2014 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 Murakami et al., 2014 for details.

METHOD DETAILS

Pattern sequence estimation

A 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 Mazzucato et al., 2015, 2019 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,” Figure 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 conditional on the pattern’s firing rates $r_i(m)$. Stochastic transitions between patterns occur according to a Markov chain with transition matrix (TPM, Figure 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 $y_i(t) = 1$ (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 model selection procedure (Engel et al., 2016). 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 Figure S2). The held-out $LL(M)$ increases with M, until reaching a plateau. We selected the number of patterns M^* for which the incremental increase $LL(M+1) - LL(M)$ had the largest drop (the point of largest curvature) before the plateau (Satopaa et al., 2011). For control, we performed model selection using an alternative method, the Bayesian Information Criterion (Mazzucato et al., 2019), 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 (Tomasi and Bro, 2006):

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

where N is the ensemble size, $\widehat{r}_i^{[k]}(m) = r_i^{[k]}(m) / \left|\left| \overrightarrow{r}^{[k]}(m) \right|\right|_2$ is the normalized emission for pattern m, and $\widehat{T}_{mn}^{[k]}$ is the normalized transition matrix $\widehat{T}_{mn}^{[k]} = T_{mn}^{[k]} / \left|\left| T^{[k]} \right|\right|_2$. Features were matched across folds using the stable matching algorithm (Gale and Shapley, 2013). 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 (Williams et al., 2018). 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 Figure 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, Figure 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 (Figure 1F) within each session were estimated from the empirical distribution of interval times where a pattern's probability was above 80%. Lowering the value of 50ms would extend the distribution of pattern dwell times toward zero (*cf.* Figure 1F) reducing the mean of such a distribution, although the characteristic tail and long transitions would remain as a hallmark of the underlying neural processes.

HMM robustness to neural population subsampling

We compared the HMM analysis of the full empirical dataset with two datasets obtained by removing specific neurons in each session. The two datasets were obtained as follows. For each session we computed the average firing rate for each neuron across all trials without taking HMM states into account, Figure S4B. Then in one case (first dataset) we removed the neuron with the highest firing rate and run the HMM analysis again, this is labeled "high" in Figure S4C. In the second case we removed the neuron with median firing rate and similarly run the HMM analysis, "median" case in Figure S4D. Examples of the outcome of the HMM fit on the example session excluding respectively the top firing and median firing neuron are shown in Figures S4C and S4D. The HMM fits to the subsampled statistics were in astounding agreement with the HMM fit to the full population underscoring the robustness of the HMM method even for small population sizes. In our statistics the average number of neurons per session was 9.9 ± 3.3. This robustness was captured by several metrics Figures S4E and S4F. The number of states individuated selected by the HMM was remarkably similar across all sessions Figure S4E, both concerning the number of states selected by our crossvalidation procedure and the number of states retained by our selection criteria, *cf.* STAR Methods 4.3. Similarly properties related to the sequence of states (i.e., number of states per sequence, average state duraction and fraction of time for each trial occupied by states with a 80% posterior probability) were all remarkably similar with no significant difference across all sessions Figure S4F.

Comparison with surrogate datasets

We compared the HMM analysis of the empirical dataset with two surrogate datasets, obtained with the following shuffled procedure (Figure 2, (Maboudi et al., 2018)). 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, packets of binned population spike counts were randomly permuted in time (column-wise swap), preserving pairwise correlations but reducing autocorrelations. Each packet consisted of 10 binned spike count vectors amounting to a total time of 50 ms as each bin was of 5 ms. 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. The TPM returns the probability for a transition between two patterns to occur (see Pattern sequence estimation), therefore a sparser TPM suggests a more robust sequence unfolding. 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 $n \rightarrow m$ transitions observed; finally each row was normalized: $T_{mn}^{(sym)} \rightarrow T_{mn}^{(sym)} / \sum_{l=1}^{N} T_{ml}$. E.g., the pattern sequence $1,2,3,1,2$ is in one-to-one correspondence to the symbolic TPM

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

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.

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

$$r_i(m) = -\frac{1}{dt}\log\left(1-\frac{\sum_{t=1}^{T}p(m,t)y_i(t)}{\sum_{t=1}^{T}p(m,t)}\right), \qquad \text{(Equation 3)}$$

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 (Figure 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 ($p_{mn} < 0.05/M$) and $S_{mn} = 1$ otherwise. For example, consider the case of 3 patterns and the following S 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 (Mazzucato et al., 2015).

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 $\overline{m}$ when $p(\overline{m},\bar{t}) \geq 0.8$. When this criterion was not met by any pattern then no pattern was assigned, *cf.* Figure 1D. The onset time of a specific pattern $\overline{m}$ was identified as the first time $\bar{t}$ where $p(\overline{m},\bar{t}) \geq 0.8$. Transitions of several patterns appeared in close proximity to specific events (Figure 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 $\mathcal{T}(\overline{m})$ for the onset times of pattern $\overline{m}$. If the average of the distribution $\mu(\mathcal{T}(\overline{m})) \in [-0.5, 0.1]sec$, we tagged the pattern $\overline{m}$ to the event. When multiple transitions matched our criteria, we selected the one with minimum inter-quartile $iqr(\mathcal{T}(\overline{m}))$. This procedure returned patterns tagged with specific actions for each trial (*cf.* Figure 4C) and tagged one or more patterns in 82% of the sessions. Wherever a pattern was tagged it appeared on average in 90% of the session's trials. We name pattern onset times $\{t_{PI}, t_{PO}, t_{WPI}\}$ respectively for the actions Poke In, Poke Out and Water Poke In.

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_{action} and the tagged pattern onset times $t_{\overline{e}} \in \{t_{PI}, t_{PO}, t_{WPI}\}$, we classified the action according to

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

When no patterns passed this criteria the action was not labeled. This procedure labeled 63% of all actions. This procedure labeled 63% of all actions (at least one pattern was tagged to actions in 82% of the sessions). Whenever a pattern was tagged to an action, the pattern appeared in 90% (on average) of the session's trials. The tagging procedure was therefore robust, and we believe that it could significantly improved in future experiments with the availability of larger populations of simultaneously recorded neurons. For each session and all tagged actions we estimated a confusion matrix of our decoding procedure (*cf.* Figure 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. In order to show, in our analysis, the statistics of non-classified actions despite tagged states being present in the trial, we performed a second analysis. We limited the statistics of sessions and trials to those where all three actions had corresponding tagged patterns. In such sessions each tagged transition could be misclassified with a different action or with no action at all. Thus, it was possible to uniformly report the statistics comparing the relative occurrence of the prediction of an action versus no-action; confusion matrices for these reduced statistics are reported in Figure S6D.

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):

$$C(m)_{ij} = \frac{1}{N_T}\sum_{a}^{N_T}\left(r_i^a(m)r_j^{a,T}(m) - \left(\frac{1}{N_T}\sum_{a}^{N_T} r_i^a(m)\right)\left(\frac{1}{N_T}\sum_{a}^{N_T} r_j^{a,T}(m)\right)\right), \qquad \text{(Equation 4)}$$

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 Equation 3. We then computed the dimensionality $d(m)$ of population activity fluctuations around pattern m as the participation ratio (Abbott et al., 2011; Mazzucato et al., 2016):

$$d(m) = \frac{Tr[C(m)]^2}{Tr\left[C(m)^2\right]} = \frac{\left(\sum_i^N \lambda_i\right)^2}{\sum_i^N \lambda_i^2}, \qquad \text{(Equation 5)}$$

where λ_i are the eigenvalues of the covariance matrix for $i = 1,\ldots,N$ neurons (Abbott et al., 2011; Mazzucato et al., 2016). 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 Equation 4, where K is the integer minor or equal to the average of $d(m)$ 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.* Figure 6B. To compute the shuffled statistics for Figure 6 and similar, we proceeded as follows. In the case of the dimensionality (Figure 6A) we created random ensambles of pattern activities by shuffling (using random permutations) neural activities within a session across all patterns and neurons. In the case of the of the pattern alignment (Figure 6B) we computed the shuffled statistics by computing canonical correlation coefficients not between top principal components of two patterns (as many as individuated by the average dimensionality described above in each session) but rather between random principal components directions. In such a way the plots highlight how top principal components of pattern correlations are aligned between two patterns against the null hypothesis of random alignment between any two principal components of the same patterns.

Firing rate modulations by stimuli and conditions

Single cell responses to tones

To estimate single neuron responses to the tone we performed a t test for responses of individual neurons before and after the two tones from the spike count vector, across trials, before and after the sensory stimulus in a window of 50 ms. We retained only trials where no other event (e.g., Poke Out) was present within a 100 ms window from the onset of the stimulus.

Decoding of condition

We sought to decode patient versus impatient trials from the time course of neural activity. We started by computing spike count vectors binning spikes of neural activity through a non-overlapping moving window of 50 ms. In each trial we considered the neural activity from 1 s before Poke In until 1.5 s after Water Poke In. We then labeled each spike count vector respectively as “patient” or “impatient” if they belonged to a trial where the animal displayed the corresponding behavior. Finally we used a neural classifier (linear or SVM) trained on all spike count vectors (Figure S6F). In the SVM case several kernels yielded similar results, displayed are results for a Gaussian kernel.

GPFA fit to neural data

A Gaussian Process Factor Analysis (GPFA (Byron et al., 2008)) represents a continuous latent space model where hidden factors underlying the dynamics smoothly unfold through time giving rise to neural activity. GPFA posits that population activity is generated by an underlying continuous and low dimensional Gaussian process. By its nature, the GPFA aims to fit a continuous latent trajectory to population activity, as opposed to the HMM’s discrete nature of sudden transitions between long-dwelling states. Although this hypothesis may theoretically provide an alternative explanation for our results, we discovered that the time course of the GPFA latent factors closely matched the HMM discrete pattern sequences.

The number of latent variables $N_{factors}$ for each session was identified by means of a 3-fold crossvalidation procedure and selected by means of choosing the point of maximum curvature in the crossvalidation curve, identically to the criterion used to identify the number of HMM states. Here we used the knee locator algorithm described in (Satopaa et al., 2011) with sensitivity parameter

$s = 1$. Once the number of latent variables was assessed for each session, a GPFA model with the corresponding number of variables was fit.

The application of the GPFA method resulted in a time series for each of the $N_{factors}$ in each trial which were underlying neural dynamics. These time series are displayed in Figure S5A for the example session considered in Figures 1B–1E. To aid visualization, for each factor (in each of the top 5 panels) we time warped the temporal dynamics in each trial to visualize it across all trials of the session. The characteristic trend of latent factors in this session shows how the onset of different actions is characterized by strong modulations of a subset of latent factors, capturing sudden changes in the neural dynamics they fit. This hints to a discrete rather than continuous feature of the latent space underlying neural dynamics. We visualized the trajectory of latent factors in their PC space Figure S5B confirming a similar signature. The marginal distribution of PC 1 and PC 2 are shown in the plots at margin.

Network model

In this section we describe the correlated variability model generating reliable sequences of metastable attractors (see Equation 1), whose dynamics is ruled by the current-based formulation of the standard rate model (Grossberg, 1969; Miller and Fumarola, 2012):

$$\tau \dot{u}_i(t) = -u_i(t) + \sum_{j=1}^{N} J^S_{ij}\phi_j(u_j(t)) + \zeta(t)\sum_{j=1}^{N} J^F_{ij}\phi_j(u_j(t)) \,. \qquad \text{(Equation 6)}$$

The firing rates are analog positive variables given by the transformation of synaptic currents to rates by the input-output transfer function $\phi_i(u_i)$. Transfer functions ϕ_i were inferred from the empirical firing rate distribution of M2 single neurons (see Inferring the transfer function from data). The parameter τ corresponds to the single neuron time constant. We set the M2 symmetric connectivity to be sparse (Mason et al., 1991; Markram et al., 1997; Holmgren et al., 2003; Thomson and Lamy, 2007; Lefort et al., 2009). Our connectivity consists of two terms, traditionally referred to as the *symmetric* term J^S_{ij} and the *asymmetric* term J^F_{ij} (Domany et al., 1995). The symmetric term reads

$$J^S_{ij} = \frac{c_{ij}A_S}{Nc}\sum_{\mu=1}^{p} f\left[\eta^\mu_i\right] g\left[\eta^\mu_j\right] \,, \qquad \text{(Equation 7)}$$

where the variable c_{ij} represents the structural connectivity of the M2 local circuit, modeled as an Erdos-Renyi graph where $c_{ij} = 1$ with probability c. The normalization constant Nc corresponds to the average number of connections to a neuron; A_S is the overall strength of the symmetric term.

For any nonzero synaptic connection (i.e., $c_{ij} = 1$), the strength of the synaptic weight is given by $\frac{A_S}{Nc}\sum_{\mu=1}^{p} f[\eta^\mu_i]g[\eta^\mu_j]$. The variables η^μ_i are distributed as $\eta^\mu_i \tilde{\approx} \phi(z^\mu_i)$, where z^μ_i are normally distributed, i.e., $z^\mu_i \tilde{\approx} N(0,1)$. The functions f and g are given by the step functions

$$f(\eta) = \begin{cases} q_f & \text{if } x_f \le \eta \\ -(1-q_f) & \text{if } \eta \le x_f \end{cases}, \quad g(\eta) = \begin{cases} q_g & \text{if } x_g \le \eta \\ -(1-q_g) & \text{if } \eta \le x_g \end{cases} \qquad \text{(Equation 8)}$$

Therefore, the pair of binary random patterns $f[\eta^\mu_i]$ and $g[\eta^\mu_i]$ are correlated. We assume that $\langle g \rangle = 0$, which constrains one of the two parameters of g. This constrain does not apply to f (i.e., $\langle f \rangle = 0$), and in our model the function f is biased toward inhibition (i.e., $\langle f \rangle < 0$). While J^S_{ij} 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 (Sompolinsky and Kanter, 1986; Kleinfeld, 1986; Herz et al., 1989; Domany et al., 1995).

The *correlated variability* term $\zeta(t)\sum_{j=1}^{N} J^F_{ij}\phi(u_j(t))$ in Equation 6 is comprised by the asymmetric matrix J^F_{ij} which is is given by

$$J^F_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} f\left(\eta^{\mu+1}_i\right) g\left(\eta^\mu_j\right), \qquad \text{(Equation 9)}$$

where the rank p of the matrix J^F_{ij} 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 $\zeta(t)$ given by

$$\tau_\zeta \dot{\zeta}(t) = -\zeta(t) + \overline{\zeta} + \sqrt{2\sigma^2_\zeta \tau_\zeta}\, x(t) \,, \qquad \text{(Equation 10)}$$

where τ_ζ, $\overline{\zeta}$ and σ^2_ζ are the timescale, mean and variance of the process, respectively. For a derivation of these parameters see the next section.

In Correlated variability is necessary to explain temporal variability in attractor networks, we compared the correlated variability model (see Equation 6) to a *private noise* model

$$\tau \dot{u}_i(t) = -u_i(t) + \sum_{j=1}^{N} J^S_{ij}\phi_j(u_j(t)) + \overline{\zeta}\sum_{j=1}^{N} J^F_{ij}\phi_j(u_j(t)) + \sqrt{2\sigma^2_p \tau}\,\chi_i(t) \,, \qquad \text{(Equation 11)}$$

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 $\overline{\zeta}$, unlike the time varying asymmetric term in Equation 6.

As a measure of the pattern retrieval (Figure 5), we used overlaps, defined as the Pearson correlation between the instantaneous firing rate and the nonlinear transformation of a given pattern $g[\overrightarrow{\eta}^l]$ (Pereira and Brunel, 2018; Gillett et al., 2020)

$$m_l(t) = \frac{Cov[g[\overrightarrow{\eta}^l]\overrightarrow{r}(t)]}{\sqrt{Var(g[\overrightarrow{\eta}^l])Var(\overrightarrow{r}(t))}} . \qquad \text{(Equation 12)}$$

Two-area mesoscale model

In this section, we show how to obtain the network model in Equation 1, starting from the two-area network in Correlated variability originates in a mesoscale feedback loop, whose dynamics is governed by

$$\tau\dot{u}_i(t) = -u_i(t) + \sum_{j=1}^{N} J_{ij}^S\phi_j(u_j(t)) + \sum_{j=1}^{N_Y} W_{ij}^{M2\leftarrow Y} r_j^Y \qquad \text{(Equation 13)}$$

$$\tau_Y \dot{r}_i^Y = -r_i^Y + \sum_{j=1}^{N} W_{ij}^{Y\leftarrow M2}\phi_j(u_j) .$$

In Figure 5E a schematic of the two-area network model is shown. The first equation describes the dynamics of N neurons in area M2, with the notations as in Equation 6. The second term represents the firing rates r_i^Y of N_Y neurons in area Y, where we assume $N_Y \ll N$. We approximate area Y dynamics as linear. The projections $W_{ij}^{Y\leftarrow M2}$ from M2 to area Y are structured similarly as M2 recurrent connections in Equation 7, i.e.,

$$W_{ij}^{Y\leftarrow M2} = \frac{A_{Y\leftarrow M2}}{N}\sum_{\mu=1}^{p-1} f(y_i^\mu)g\left(\eta_j^\mu\right), \qquad \text{(Equation 14)}$$

although there are two important differences. First, since the number of neurons in M2 and area Y are N and N_Y respectively, then $W_{ij}^{Y\leftarrow M2}$ is a rectangular matrix with dimensions $N_Y \times N$. Second, projections from Y to M2 are dense. The right vectors in the outer product inside the sum in Equation 14, $g(\eta^\mu)$, are the same vectors as the right vectors in the Equation 7. The left vectors in the outer product $g(y_i^\mu)$ are different than $g(\eta^\mu)$, but have the same statistics than η_i^μ, i.e., $y_i^\mu \tilde{=} \phi(z_i^\mu)$, where z_i^μ are normally distributed $z_i^\mu \tilde{=} N(0,1)$. Therefore, when M2 activity is in the µth attractor, since the µth overlap is order one, the activity is mostly projected along the $g(y_i^\mu)$ direction in area Y. The parameter $A_{Y\leftarrow M2}$ corresponds to the overall strength of the $M2 \rightarrow Y$ projections.

We assume the activity of area Y is fast with respect to M2 ($\tau_Y \ll \tau$), replacing the second dynamical equation in Equation 13 by its steady state

$$r_i^Y = \sum_{j=1}^{N} W_{ij}^{Y\leftarrow M2}\phi(u_j) . \qquad \text{(Equation 15)}$$

Similarly to the M2 → Y projections, the feedback projections Y → M2 are given by

$$W_{ij}^{M2\leftarrow Y} = s_{ij}^{M2\leftarrow Y}(t)\frac{A_{M2\leftarrow Y}}{N_Y}\sum_{\mu=1}^{p-1} f\left(\eta_i^{\mu+1}\right)g\left(y_j^\mu\right) . \qquad \text{(Equation 16)}$$

The variable $s_{ij}^{M2\leftarrow Y}(t)$ represents the synaptic efficacies from Y to $M2$, which in our model fluctuate in time. The parameter $A_{M2\leftarrow Y}$ corresponds to the overall strength of the $Y \rightarrow M2$ projections. Fluctuations in the synaptic efficacy are pervasive. They can be the product of several different cellular mechanisms as for example short term depression (Tsodyks and Markram, 1997) or variability in the synaptic vesicle release (Dobrunz and Stevens, 1997). In this work, we consider a simple model for capturing the temporal fluctuations in the synaptic efficacy given by a noisy linear dynamics below

$$\dot{s}_{ij}^{M2\leftarrow Y} = \frac{1 - s_{ij}^{M2\leftarrow Y}(t)}{\tau_s} + \sqrt{\frac{2\sigma_{M2\leftarrow Y}}{\tau_s}}\xi_j^{M2\leftarrow Y}(t). \qquad \text{(Equation 17)}$$

Here τ_s corresponds to the time-scale of the fluctuations in the synaptic efficacy, $\sigma_{M2\leftarrow Y}^2$ is the variance of these fluctuations, and $\xi_j^{M2\leftarrow Y}(t)$ is a Gaussian random variable with mean zero and variance one. In our model, changes on synaptic efficacy of the Y→ M2 projections depend on fluctuations of pre-synaptic neurons in area Y given by the variable $\xi_j^{M2\leftarrow Y}(t)$.

The input current to M2 due to the feedback loop between M2 and area Y is approximately

$$\sum_{l=1}^{N_Y} W_{il}^{Y\rightarrow M2} r_l^Y = \sum_{j=1}^{N} \sum_{l=1}^{N_Y} W_{il}^{M2\leftarrow Y} W_{lj}^{Y\leftarrow M2} \phi(u_j) \qquad \text{(Equation 18)}$$

$$= \frac{1}{N}\left(\overline{\zeta} + \frac{\sigma p}{\sqrt{N_Y}}\omega(t)\right) \sum_{j=1}^{N} \sum_{\mu=1}^{p} f\left(\eta_i^{\mu+1}\right) g\left(\eta_j^{\mu}\right) \phi(u_j).$$

Here, we used the fact that $\frac{1}{N_Y}\sum_{l=1}^{N_Y} f(y_l^{\mu}) g(y_l^{\mu'}) s_{il}^{M2\leftarrow Y}(t)$ has mean $\overline{\zeta}\delta_{\mu,\mu'}$ 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 Equation 17. The variable $\omega(t)$ represents the normalized fluctuations with mean zero and unit variance, and a finite autocovariance time-scale. Therefore, the matrix J_{ij}^F in Equation 1 corresponds to the effective connectivity arising from the feedback loop between M2 and area Y given by

$$J_{ij}^F = \frac{\zeta(t)}{N} \sum_{\mu=1}^{p} f\left(\eta_i^{\mu+1}\right) g\left(\eta_j^{\mu}\right), \qquad \text{(Equation 19)}$$

which has rank $p \ll N$. Assuming $p \sim \mathcal{O}(\sqrt{N_Y}) \ll N$, the fluctuations in Equation 18 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 Equation 10. Notice that τ_ζ is the effective time-scale of the temporal fluctuations in the sum over Y neurons in Equation 18. Its mean $\overline{\zeta}$ and variance σ_ζ^2 control, respectively, the strength and the variability of the effective asymmetric 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. Network simulations of our two-area model confirm our mathematical results (see Figure 5F).

Inferring the transfer function from data

For inferring the input-output transfer function from *in vivo* recordings we adapted a method proposed in (Lim et al., 2015) 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 (Lim et al., 2015), 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 Figure S7). Similarly to (Pereira and Brunel, 2018), for each recorded unit we fit this curve with a sigmoidal function

$$\phi(u) = \frac{R_0}{1 + e^{-\beta(u-h_0)}}. \qquad \text{(Equation 20)}$$

If input currents produced firing rates in Equation 20 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 $\{(R_{max}^{(i)}, R_0^{(i)}, \beta^{(i)}, h_0^{(i)})\}_{i=1}^{328}$, one from each recorded unit (Figure 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.

Network simulations

For the network simulations of the correlated variability model in Figs. Figures 5 and 6, the parameter values used are listed in Table 1. 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 Equation 7) 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 Figures S7C and S7D the parameter are the same as in Table Table 1 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.

HMM fit to network model simulations

To verify our framework and the properties of our model, we fit a HMM model to the network simulations of the model and recomputed the same plots of Figures 5 and 6 reported in Figure S9. Neural activities in the model are of a continuous nature (rate based simulations) therefore we fit a Gaussian HMM as follows. We run simulations of the model by matching the statistical properties of empirical data (see previous section). For each simulated session we sampled a number of neurons equal to the one in the recordings. We then convolved neural activities with a Gaussian filter of short duration (sd 60 ms). We experienced this to be necessary to obtain a reliable convergence of the HMM. We selected the number of states in the model and fit the simulated neural population activities for each session with the HMM by means of a expectation-maximization algorithm. This fit returned the identity of the state generating the neural activity. On these statistics we then run the analysis of Figures 5 and 6 to obtain the metrics shown in Figure S9.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data analysis 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 (Erlich et al., 2011; Guo et al., 2014). All summary statistics are mean ± SD across 33 sessions, unless otherwise stated.