Stability of working memory in continuous attractor networks under the control of short-term plasticity

Alexander Seeholzer[1], Moritz Deger[1,2], Wulfram Gerstner[1*]

1 School of Computer and Communication Sciences and School of Life Sciences, Brain Mind Institute, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
2 Institute for Zoology, Faculty of Mathematics and Natural Sciences, University of Cologne, Cologne, Germany

* wulfram.gerstner@epfl.ch

Abstract

Continuous attractor models of working-memory store continuous-valued information in continuous state-spaces, but are sensitive to noise processes that degrade memory retention. Short-term synaptic plasticity of recurrent synapses has previously been shown to affect continuous attractor systems: short-term facilitation can stabilize memory retention, while short-term depression possibly increases continuous attractor volatility. However, it currently remains unclear to which degree these two short-term plasticity mechanisms interact, what their combined quantitative effect on working memory stability is, and whether these effects persist in neuronal networks with spike-based transmission. Here, we present a comprehensive description of the effects of short-term plasticity on noise-induced memory degradation in one-dimensional continuous attractor models. Our theoretical description, applicable to spiking and rate-based models alike, accurately describes the slow dynamics of stored memory positions in separate processes of diffusion due to spiking variability and drift due to sparse connectivity and neuronal heterogeneity. We find that facilitation decreases both diffusion and directed drifts, while short-term depression tends to increase both. Using mutual information, we evaluate the combined impact of short-term facilitation and depression on the ability of networks to retain stable working memory. Finally, our theory establishes links to experiments: we are able to predict the sensitivity of continuous working memory to distractor inputs and place constraints on network and synapse properties necessary to implement stable working memory.

Author summary

The ability to transiently memorize positions in the visual field is crucial for behavior. Models and experiments have shown that such memories can be maintained in networks of cortical neurons with a continuum of possible activity states, that reflects the continuum of positions in the environment. However, the accuracy of positions stored in such networks will degrade over time due to the noisiness of neuronal signaling and imperfections of the biological substrate. Previous work in simplified models has shown that synaptic short-term plasticity could stabilize this degradation by dynamically up- or down-regulating the strength of synaptic connections, thereby "pinning down" memorized positions. Here, we present a general theory that accurately predicts the extent of this "pinning down" by short-term plasticity in a broad class of biologically

plausible models, thereby untangling the interplay of varying biological sources of noise with short-term plasticity. Importantly, our work provides a direct and novel theoretical link from the microscopic substrate of working memory – neurons and synaptic connections – to observable behavioral correlates. This allows us to constrain properties of cortical networks that are currently hard to assess experimentally, which we hope will help guide future theoretical and experimental work.

Introduction

Information about past environmental stimuli can be stored and retrieved seconds later from working memory [1, 2]. Strikingly, this transient storage is achieved for timescales of seconds with neurons and synapse transmission operating mostly on time scales of tens of milliseconds and shorter [3]. An influential hypothesis of neuroscience is that working memory emerges from recurrently connected cortical neuronal networks: memories are retained by self-generating cortical activity through positive feedback [4–7], thereby bridging the time scales from milliseconds (neuronal dynamics) to seconds (behavior).

Sensory stimuli are often embedded in a physical continuum: for example, positions of objects in the visual field, frequencies of auditory stimuli, or the intensity and position of somatosensory stimuli on the body, all have continuously varying values. Ideally, the organization of cortical working memory circuits should reflect the continuous nature of sensory information [3]. A class of cortical working memory models able to store continuously structured information is that of *continuous attractors*, characterized by a continuum of meta-stable states, which can be used to retain memories over delay periods much longer than those of the single network constituents [8]. Continuous attractors were originally proposed as theoretical models for cortical working memory [9–11], path integration [12–14], and other cortical functions [15–17] (see e.g. [3, 18–21] for recent reviews), but recent experimental evidence for continuous attractor dynamics was indeed found in cortical networks [22], the limbic system [18, 23], and one-dimensional ring-attractors in the fly responsible for path integration and self-orientation [24, 25].

Continuous attractor models have been successfully employed in the context of visuospatial working memory to explain behavioral performance [26–29], to predict the effects of neuromodulation [30, 31], or the implications of cognitive impairment [32, 33]. However, when embedded in more realistic scenarios, the continuum of states quickly breaks down, since encoded memories are vulnerable to noise and heterogeneities that break, transiently or permanently, the crucial symmetry necessary for continuous attractors [10, 11, 13, 16, 34–40]. For example, the stochasticity of neuronal spiking ("fast noise") leads to transient asymmetries that randomly displace encoded memories along the continuum of states [10, 11, 35, 37, 39, 40], leading, averaged over many trials, to *diffusion* of encoded information. More drastically, introducing fixed asymmetries ("frozen noise") due to network heterogeneities causes a *directed drift* of memories and a collapse of the continuum of attractive states to a set of discrete states. Examples of heterogeneities in biological scenarios include the sparsity of recurrent connections [13, 36], or randomness in neuronal parameters [36] and values of recurrent weights [16, 34, 38]. Since both (fast) noise and heterogeneities are expected in cortical settings, the feasibility of continuous attractors as computational systems of the brain has been called into question [3, 6, 41].

Short-term plasticity of recurrent synaptic connections has been shown to influence the susceptibility of continuous attractor networks to both noise and heterogeneities. In particular, short-term depression has been observed to have strong effects on directed displacement of attractor states in rate models [42, 43], although no such effect was

apparent in a spiking network implementation [44]. Short-term facilitation, on the other hand, has been proposed as a stabilizing mechanisms that could increase the retention time of memories in continuous attractor networks with noise-free rate neurons [38]. In simulations of a continuous attractor implemented with spiking neurons, facilitation with a fixed set of parameters was reported to cause slow drift [45, 46] and a reduced amount of diffusion [46]. However, despite the large number of existing studies, several fundamental questions remain unanswered. What are the quantitative effects of short-term facilitation in more complex neuronal models and across facilitation parameters? Does short-term depression influence the strength of diffusion and drift, and how does it interplay with facilitation? Do phenomena reported in rate networks persist in spiking networks? Finally, can a single theory be used to predict all of the effects observed in simulations?

Here, we present a comprehensive description of the effects of short-term facilitation and depression on noise-induced displacement of one-dimensional continuous attractor models. Extending earlier theories for diffusion [39, 40] and drift [38], we derive predictions of the amount of diffusion and drift in ring-attractor models with short-term plasticity. Our theory generalizes to a large class of neuron models, given their input-output relation, and to various biologically plausible sources of heterogeneity. The theoretical predictions are validated against a set of ring-attractor networks realized with spiking neurons for varying parameters of short-term facilitation and depression. We find that facilitation and depression play antagonistic roles: facilitation can decrease *both diffusion and drift* while depression *increases both*, which we show to have profound impact on the retention of memories, as measured by mutual information. Importantly, our theory is, to a large degree, independent of the microscopic network configurations, which allows relating it to experimentally observable quantities. We apply this insight to show that our theory can predict the sensitivity of networks with short-term plasticity to distractor stimuli. Finally, we demonstrate a second possible link to experiments by placing theoretical bounds on combinations of network and synapse properties that will yield stable working memory (as predicted by our model) under the presence of realistic biological variability.

Results

We investigated, in theory and simulations, the effects of short-term synaptic plasticity (STP) on the dynamics of ring-attractor models consisting of N excitatory neurons with distance-dependent and symmetric excitation, and global (uniform) inhibition provided by a population of inhibitory neurons (Fig. 1A). For simplicity, we describe neurons in terms of firing rates, but our theory can be mapped to more complex neurons with spiking dynamics. An excitatory neuron i with $0 \leq i < N$ is assigned an angular position $\theta_i = \frac{2\pi}{N} i - \pi \in [-\pi, \pi)$, where we identify the bounds of the interval to form a ring topology (Fig. 1A). The firing rate ϕ_i (in units of Hz) for each excitatory neuron i ($0 \leq i < N-1$) is given as a function of the neuronal input:

$$\phi_i(t) = F(J_i(t) + J_{\text{inh}}). \tag{1}$$

Here, the input-output relation F relates the dimensionless excitatory J_i and inhibitory J_{inh} inputs of neuron i to its firing rate. This represents a rate-based simplification of the possibly complex underlying neuronal dynamics [47]. We assume that the excitatory input $J_i(t)$ to neuron i at time t is given by a sum over all presynaptic neurons

$$J_i(t) = \sum_{j=0}^{N-1} w_{ij} s_j(t), \tag{2}$$

where $w_{ij}s_j(t)$ describes the total activation of synaptic input from the presynaptic neuron j onto neurons i. The maximal strength w_{ij} of recurrent excitatory-to-excitatory connections is chosen to be local in the angular arrangement of neurons, such that connections are strongest to nearby excitatory neurons (Fig. 1A, red lines). The momentary input depends also on the synaptic activation variables s_j, to be defined below. Finally, connections to and from inhibitory neurons are assumed to be uniform and global (all-to-all) (Fig. 1A, blue lines), thereby providing non-selective inhibitory input J_{inh} to excitatory neurons.

Fig 1. Drift and diffusion in ring-attractor models with short-term plasticity. A Excitatory (E) neurons (red circles) are distributed on a ring with coordinates in $[-\pi, \pi]$. Excitatory-to-excitatory (E-E) connections (red lines) are distance-dependent, symmetric, and subject to short-term plasticity (facilitation and depression, see Eq. (3)). Inhibitory (I) neurons (blue circles) project to all E and I neurons (blue lines) and receive connection from all E neurons (gray lines). Only outgoing connections from shaded neurons are displayed. In spiking simulations, neurons additionally receive excitatory input with spikes generated by homogeneous Poisson processes. **B1** Example simulation: E neurons fire asynchronously and irregularly at low rates until (dotted line) a subgroup of E neurons is stimulated (external cue), causing them to spike at elevated rates (red dots, input was centered at 0, starting at $t = 2s$ for $1s$). During and after (dashed line) the stimulus, a bump state of elevated activity forms and sustains itself after the external cue is turned off. The spatial center of the population activity is estimated from the momentary firing rates (red line, plotted from $t = 2.5s$ onward). Inset: Activity profile in the bump state, centered at 0. **B2** Center positions of 20 repeated spiking simulations for 10 different initial cue positions each for a network with short-term depression ($U = 1, \tau_x = 150ms$). Random E-E connections (with connection probability $p = 0.5$) lead to directed drift in addition to diffusion. Right: Normalized histogram (200 bins) of final positions at time $t = 13.5$. **C** Illustration of quantities used in theoretic calculations. Neurons in the bump fire at rates $\phi_{0,i}$ (dashed black line, compare to B1, inset) due to the steady-state synaptic input $J_{0,i}$ (blue line). Movement of the bump center causes a change of the synaptic input $\frac{dJ_{0,i}}{d\varphi}$ (orange line). **D** Diffusion along the attractor manifold is calculated (see Eq. (5)) as a weighted sum of the neuronal firing rates in the bump state (dashed black line). Spiking noise (red line) is illustrated as a random deviation from the mean rate with variance proportional to the rate. The symmetric weighting factors (blue lines show $\frac{C_i^2}{S^2}\left(\frac{dJ_{0,i}}{d\varphi}\right)^2$ for varying U) are non-zero at the flanks of the firing rate profile. Stronger short-term depression and weaker facilitation increase the magnitude of weighting factors. **E** Deterministic drift is calculated as a weighted sum (see Eq. (7)) of systematic deviations of firing rates from the bump state (frozen noise): a large positive firing rate deviation in the left flank (red line) will cause movement of the center position to the left (red arrow) because the weighting factors (blue lines show $\frac{C_i}{S}\frac{dJ_{0,i}}{d\varphi}$ for varying U) are asymmetric.

As a model of STP, we assume that excitatory-to-excitatory connections are subject to short-term facilitation and depression, which we implemented using a widely adopted model of short-term synaptic plasticity [48]. The *outgoing synaptic activations* s_j of

neuron j are modeled by the following system of ordinary differential equations:

$$\begin{aligned} \dot{s}_j &= -\frac{s_j}{\tau_s} + u_j x_j \phi_j, \\ \dot{u}_j &= -\frac{u_j - U}{\tau_u} + U(1-u_j)\phi_j, \\ \dot{x}_j &= -\frac{x_j - 1}{\tau_x} - u_j x_j \phi_j. \end{aligned} \tag{3}$$

The timescale of recovery τ_x is the main parameter of depression. While the recovery from facilitation is controlled by the timescale τ_u, the parameter $0 < U \leq 1$ controls the baseline strength of unfacilitated synapses as well as the timescale of their strengthening. For fixed τ_u, we consider smaller values of U to lead to a "stronger" effect of facilitation, and take $U = 1$ as the limit of non-facilitating synapses.

As a reference implementation of this model, we simulated networks of spiking conductance-based leaky-integrate-and-fire (LIF) neurons with (spike-based) short-term plastic synaptic transmission (Fig. 1B1, see *Spiking network model* in Materials and Methods for details). For these networks, under the assumption that neurons fire with Poisson statistics and the network is in a stationary state, neuronal firing can be approximated by the input-output relation F of Eq. (1) [49, 50] (see *Firing rate approximation* in Materials and Methods), which allows us to map the network into the general framework of Eqs. (1)-(2). In the stationary state, synaptic depression will lead to a saturation of the synaptic activation variables s_j at a constant value as firing rates increase. This nonlinear behavior enables spiking networks to implement bi-stable attractor dynamics with relatively low firing rates [45, 51] similar to saturating NMDA synapses [11, 46]. Since we found that without depression (for $\tau_x \to 0$) the bump state was not stable at low firing rates (in agreement with [51]), we always keep the depression timescale τ_x at positive values.

Particular care was taken to ensure that networks display nearly identical bump shapes (similar to Fig. 1B1, inset; see also S1 Fig), which required the re-tuning of network parameters (recurrent conductance parameters and the width of distance-dependent connections; see *Optimization of network parameters* in Materials and Methods) for each combination of the STP parameters above.

Spiking simulations generally show a bi-stability between a *non-selective* state and a *bump* state. In the non-selective state, all excitatory neurons emit action potentials asynchronously and irregularly at roughly identical and low firing rates (Fig. 1B1, left of dotted line). The bump state can be evoked by stimulating excitatory neurons localized around a given position by additional external input (Fig. 1B1, red dots). After the external cue is turned off, a self-sustained firing rate profile ("bump") emerges (Fig. 1B1, right of dashed line, and inset) that persists until the network state is again changed by external input. For example, a short and strong uniform excitatory input to all excitatory neurons causes a transient increase in inhibitory feedback that is strong enough to return the network to the uniform state [11].

During the bump state, fast fluctuations in the firing of single neurons transiently break the perfect symmetry of the firing rate profile and introduce small random displacements along the attractor manifold, which become apparent as a random walk of the center position. If the simulation is repeated for several trials, the bump has the same shape in each trial, but information on the center position is lost in a *diffusion*-like process. We additionally included varying levels of biologically plausible sources of heterogeneity (frozen noise) in our networks: *random connectivity* between excitatory neurons (E-E) and heterogeneity of the single neuron properties of the excitatory population [36], realized as a random distribution of leak reversal potentials. Heterogeneities makes the bump *drift* away from its initial position in a directed

manner. For example, the bump position in the randomly connected ($p = 0.5$) network of Fig. 1B1 shows a clear upwards drift towards center positions around 0. Repeated simulations of the same attractor network with bumps initialized at different positions provide a more detailed picture of the combined drift and diffusion dynamics: bump center trajectories systematically are biased towards a few stable fixed points (Fig. 1B2) around which they are distributed for longer simulation times (histogram in Fig. 1B2, $t = 13.5s$). The theory developed in this paper aims at analyzing the above phenomena of drift and diffusion of the bump center.

Theory of diffusion and drift with short-term plasticity

To untangle the observed interplay between diffusion and drift and investigate the effects of short-term plasticity, we derived a theory that reduces the microscopic network dynamics to a simple one-dimensional stochastic differential equation for the bump state. The theory yields analytical expressions for diffusion coefficients and drift fields, that depend on short-term plasticity parameters, the shape of the firing rate profile of the bump, as well as the neuron model chosen to implement the attractor.

First, we *assume* that the system of Eqs. (3) together with the network Eqs. (1)-(2) has a 1-dimensional manifold of meta-stable states, i.e. the network is a ring-attractor network as described in the introduction. This entails, that the network dynamics permit the existence of a family of solutions that can be described as a self-sustained and symmetric bump of firing rates $\phi_{0,i}(\varphi) = F\left(J_{0,i}(\varphi)\right)$ with corresponding inputs $J_{0,i}(\varphi)$ (for $0 \leq i < N$). Importantly, the center φ of the bump can be located at any arbitrary position $\varphi \in \left\{ \frac{j}{N} 2\pi - \pi \middle| 0 \leq j < N \right\}$. For example, if $\phi_{0,i}(0)$ is a solution with input $J_{0,i}(0)$, then $\phi_{0,i+1}(\frac{2\pi}{N})$ is also a solution with input $J_{0,i+1}(\frac{2\pi}{N})$. This solution is illustrated in Fig. 1C for a bump centered at $\varphi = 0$. Second, we assume that the number N of excitatory neurons is large ($N \to \infty$), such that we can think of the possible positions φ as a continuum. Our final assumption is that neuronal firing is noisy, with spike counts distributed as Poisson processes, and that we are able to replace the shot-noise of Poisson spiking by white Gaussian noise with the same mean and autocorrelation (see *Diffusion* in Materials and Methods, and Discussion). Under these assumptions, we are able to reduce the network dynamics to a **one-dimensional Langevin equation**, describing the dynamics of the center $\varphi(t)$ of the firing rate profile (see *Analysis of drift & diffusion with STP* in Materials and Methods):

$$\dot{\varphi} = \sqrt{B}\eta(t) + A(\varphi). \tag{4}$$

Here, $\eta(t)$ is white Gaussian noise with zero mean and correlation function $\langle \eta(t), \eta(t') \rangle = \delta(t - t')$.

The first term is diffusion characterized by a ***diffusion strength*** B[1], which describes the random displacement of bump center positions due to fluctuations in neuronal firing. For $A(\varphi) = 0$ this term causes diffusive displacement of the center $\varphi(t)$ from its initial position $\varphi(t_0)$, with a mean (over realizations) squared displacement of positions $\left\langle [\varphi(t) - \varphi(t_0)]^2 \right\rangle = B \cdot (t - t_0)$ that, during an initial phase, increases linearly with time [14, 52, 53], before saturating due to the circular domain of possible center positions [39]. Our theory shows (see *Diffusion* in Materials and Methods) that the coefficient B can be calculated as a weighted sum over the neuronal firing rates (Fig. 1D)

$$B = \sum_i \left(\frac{C_i}{S}\right)^2 \left(\frac{dJ_{0,i}}{d\varphi}\right)^2 \phi_{0,i}, \tag{5}$$

[1] In Brownian motion, the *diffusion constant* is usually defined as $D = B/2$.

where $\frac{dJ_{0,i}}{d\varphi}$ is the change of the input to neuron i under shifts of the center position (Fig. 1C, orange line), and S is a normalizing constant.

The analytical factors C_i express the spatial dependence of the diffusion coefficient on the short-term plasticity parameters through

$$C_i = \frac{U\left(1 + 2\tau_u\phi_{0,i} + U\tau_u^2\phi_{0,i}^2\right)}{\left(1 + U\phi_{0,i}\left(\tau_u + \tau_x\right) + U\tau_u\tau_x\phi_{0,i}^2\right)^2}. \tag{6}$$

The dependence of the single summands in Eq. (5) on short-term plasticity parameters is visualized in Fig. 1D, where we see that: a) due to the squared spatial derivative $\frac{dJ_{0,i}}{d\varphi}$ of the bump shape and the squared factors C_i/S, the important contributions to the sum arise primarily from the flanks of the bump; b) for a fixed bump shape, summands increase with stronger short-term depression (larger τ_x) and decrease with stronger short-term facilitation (smaller U, larger τ_u). This result extends the approach of [39] to synaptic dynamics with short-term plasticity: the limiting case of no facilitation and depression ($U = 1$, $\tau_x = 0ms$) simplifies the normalization factor S considerably, leaves $C_i = 1$, and recovers the result for diffusion the stated there [39, Eq. S18].

The second term in Eq. (5) is the ***drift field*** $A(\varphi)$, which describes deterministic drifts due to the inclusion of heterogeneities. For heterogeneity caused by variations in neuronal reversal potentials and random network connectivity, we calculate (see *Frozen noise* in Materials and Methods) systematic deviations $\Delta\phi_i(\varphi)$ of the single neuronal firing rates from the steady-state bump shape that depend on the current position φ of the bump center. In *Drift* in Materials and Methods, we show that the drift field is then given by a weighted sum over the firing rate deviations:

$$A(\varphi) = \sum_i \frac{C_i}{S}\frac{dJ_{0,i}}{d\varphi}\Delta\phi_i(\varphi), \tag{7}$$

with weighing factors depending on the spatial derivative of the bump shape $\frac{dJ_{0,i}}{d\varphi}$ and the parameters of the synaptic dynamics through the same factors C_i/S. This is illustrated in Fig. 1E: in contrast to Eq. (5) summands are now asymmetric with respect to the bump center, since the spatial derivative is not squared.

Prediction of continuous attractor dynamics with short-term plasticity

To demonstrate the accuracy of our theory, we chose random connectivity as a first source of frozen variability. Random connectivity was realized in simulations by retaining only a random fraction $0 < p \leq 1$ (connection probability) of excitatory-to-excitatory (EE) connections. The uniform connections from and to inhibitory neurons are taken as all-to-all, since the effects of making these random and sparse would have only indirect effects on the dynamics of the bump center positions.

Our theory accurately predicts the drift-fields $A(\varphi)$ (see Eq. (7)) induced by frozen variability in networks with short-term plasticity (Fig. 2). Briefly, for each neuron $0 \leq i < N$, we treat each realization of frozen variability as a perturbation Δ_i around the perfectly symmetric system and use an expansion to first order of the input-output relation F to calculate the resulting changes in firing rates (see *Frozen noise* for details):

$$\Delta\phi_i(\varphi) = \frac{dF}{d\Delta_i}\Delta_i. \tag{8}$$

The resulting terms are then used in Eq. (7) to predict the magnitude of the drift field $A(\varphi)$ for any center position φ, which will, importantly, depend on STP parameters.

The same approach can be used to predict drift fields induced by heterogeneous single neuron parameters [36] (see next sections) and additive noise on the E-E connection weights [16, 38].

Fig 2. Drift field predictions for varying short-term facilitation. All networks have the same instantiation of random connectivity ($p = 0.5$), similar to Fig. 1B1. **A** Centers of excitatory population activity for 50 repetitions of $13.5s$ delay activity, for 20 different positions of initial cues (cue is turned off at $t = 0$) colored by position of the cues. Left: no facilitation ($U = 1$). Right: with facilitation ($U = 0.1$). **B** Drift field as a function of the bump position. The theoretical prediction (blue line, see Eq. (7)) of the drift field is compared to velocity estimations along the trajectories shown in A, colored by the line they were estimated from. The thick black line shows the binned mean of data points in 60 bins. For comparison, the predicted drift field for $U = 0.1$ is plotted (thin dashed line). Left: no facilitation ($U = 1$), for comparison the theoretical prediction for the case $U = 0.01$ is plotted as a dashed line. Right: with facilitation ($U = 0.01$). **C** Trajectories under the same conditions as in A, but obtained by forward-integrating the one-dimensional Langevin equation, Eq. (4). **D** Normalized histograms of final positions at time $t = 13.5$ for data from spiking simulations (gray areas, data from A) and forward solutions of the Langevin equations (blue areas, data from C). Other STP parameters were: $\tau_u = 650ms, \tau_x = 150ms$.

We first simulated spiking networks with only short-term depression and without facilitation (Fig. 2A, left, same network as in Fig. 1B1), for one instantiation of random random ($p = 0.5$) connectivity. Numerical estimates of the drift in spiking simulations (by measuring the displacement of bumps over time as a function of their position, see *Spiking simulations* in Materials and Methods for details) yielded drift-fields in good agreement with the theoretical prediction (Fig. 2B, left). At points where the drift field prediction crosses from positive to negative values (e.g. Fig. 2B, left, $\varphi = \frac{\pi}{2}$), we expect stable fixed points of the center position dynamics in agreement with simulation results, which show trajectories converging to these points. Similarly, unstable fixed points (negative-to-positive crossings) can be seen to lead to a separation of trajectories (e.g. Fig. 2A, left, $\varphi = -\frac{\pi}{2}$). In regions where the positional drifts are predicted to lie close to zero (e.g. Fig. 2A, left $\varphi = 0$) the effects of diffusive dynamics are more pronounced. Finally, numerical integration of the full 1-dimensional Langevin equation Eq. (4) with coefficients predicted by Eqs. (5)-(7), produces trajectories with dynamics very similar to the full spiking network (Fig. 2C, left). When comparing the center positions after $13.5s$ of delay activity between the full spiking simulation and the simple 1-dimensional Langevin system, we found very similar distributions of final positions (Fig. 2D, left, compare to Fig. 1B1, histogram). Our theory thus produces an accurate approximation of the dynamics of center positions in networks of spiking neurons with STP, thereby reducing the complex dynamics of the whole network to a simple equation. It should be noted that, in regions with strong drift or steep negative-to-positive crossings, the numerically estimated drift-fields deviate from the theory due to under-sampling of these regions as trajectories move quickly through them, yielding fewer data points. In *Short-term plasticity controls drift* we show that for stronger heterogeneities the theory tends to generally over-predict drift-fields.

Introducing strong short-term facilitation ($U = 0.1$) reduces the predicted drift fields (Fig. 2B, left, dashed line), which resemble a scaled-down version of the drift-field for the unfacilitated case. We confirmed this theoretical prediction by simulations including facilitation (Fig. 2A, right): the resulting drift fields show significant reduction of speeds (Fig. 2B, right) while zero crossings remained similar to the unfacilitated network, similar to the results in [38]. Theoretical predictions of the drift fields with bump shapes extracted from these simulations again show an accurate prediction of the

dynamics (Fig. 2B, right). Thus, as before, forward integrating the simple 1-dimensional Langevin-dynamics yields trajectories (Fig. 2C, right) highly similar to those of the full spiking network, with closely matching distributions of final positions (Fig. 2D, right), indicative of a matching strength of diffusion. In summary, our theory predicts the effects of STP on the joint dynamics of diffusion and drift due to network heterogeneities, which we will show in detail in the next sections.

Short-term plasticity controls diffusion

To isolate the effects of STP on diffusion, we simulated networks *without frozen noise* for various STP parameters. For each combination of parameters, we simulated 1000 repetitions of $13.5s$ delay activity (after cue offset) distributed across 20 uniformly spaced initial cue positions (see Fig. 3A for an example). From these simulations, the strength of diffusion was estimated by measuring the growth of variance (over repetitions) of the distance of the center position from its initial position as a function of time (see *Spiking simulations* in Materials and Methods for details). For all parameters considered, this growth was well fit by a linear function (e.g. Fig. 3A, inset), the slope of which we compared to the theoretical prediction obtained from the diffusion strength B (Eq. (5)).

Fig 3. Diffusion on continuous attractors is controlled by short-term plasticity. A Center positions of 20 repeated simulations of the reference network ($U = 1, \tau_x = 150$ms) for 10 different initial cue positions each. Inset: Estimated variance of deviations of center positions $\varphi(t)$ from their positions $\varphi(0.5)$ at $t = 0.5s$ (purple) as a function of time ($\langle[\varphi(t) - \varphi(0.5)]^2\rangle$), together with linear fit (dashed line). The slope of the dashed line yields an estimate of B (Eq. (5)). **B,C** Diffusion strengths estimated from simulations (dots, error bars show 95% confidence interval, estimated by bootstrapping) compared to theory. Dashed lines show theoretical prediction using firing rates measured from the reference network ($U = 1, \tau_x = 150ms$), while crosses are theoretical estimates using firing rates measured for each set of STP parameters separately (crosses). **B** Diffusion strength as a function of facilitation parameter U. Inset shows zoom of region indicated in the dashed area in the lower left. Increasing the facilitation time constant $\tau_u = 650ms$ (blue) to $\tau_u = 1s$ (orange) affects diffusion only slightly. In panels A and B, the depression time constant is $\tau_x = 150ms$. **C** Diffusion strength as a function of depression time constant τ_x. Results for three different values of U are shown (note the change in scale). Colors indicate the two different values for the facilitation time constant also used in panel B.

We find that facilitation and depression control the amount of diffusion along the attractor manifold in an antagonistic fashion (Fig. 3B,C). First, increasing facilitation by lowering the facilitation parameter U from its baseline $U = 1$ (no facilitation) towards $U = 0$, while keeping the depression time constant $\tau_x = 150ms$ fixed, decreases the measured diffusion strength over an order of magnitude (Fig. 3B, dots). On the other hand, increasing the facilitation time constant τ_u from $\tau_u = 650ms$ to $\tau_u = 1000ms$ (Fig. 3B, orange and blue dots, respectively) only slightly reduces diffusion. Our theory further predicts that increasing the facilitation time constants above $\tau_u = 1s$ will not lead to large reductions in the magnitude of diffusion (see S2 Fig). Second, we find that increasing the depression time constant τ_x for fixed U, thereby slowing down recovery from depression, leads to an increase of the measured diffusion (Fig. 3C). More precisely, increasing the depression time constant from $\tau_x = 120ms$ to $\tau_x = 200ms$ leads only to slight increases in diffusion for strong facilitation ($U = 0.1$), but to a much larger increase for weak facilitation ($U = 0.8$).

For a comparison of these simulations with our theory, we used two different approaches. First, we estimated the diffusion strength by using the precise shape of the stable firing rate profile extracted separately for each network with different sets of parameters. This first comparison with simulations confirms that the theory closely describes the dependence of diffusion on short-term plasticity for each parameter set (Fig. 3B, crosses). The observed effects could arise directly from changes in STP parameters for a fixed bump shape, or indirectly since STP parameters also influence the shape of the bump. To separate such direct and indirect effects, we used for a second comparison a theory with fixed bump shape, i.e. the bump shape measured in a "reference network" ($U = 1$, $\tau_x = 150ms$) and extrapolated curves by changing only STP parameters in Eq. (5). This leads to very similar predictions (Fig. 3B, dashed lines) and supports the following conclusions: a) the diffusion to be expected in attractor networks with similar observable quantities (mainly, the bump shape) depends only on the short-term plasticity parameters; b) the bump shapes in the family of networks we have investigated are sufficiently similar to be approximated by measurement in a single reference network. It should be noted that the theory tends to slightly over-estimate the amount of diffusion, especially for small facilitation U (see Fig. 3B, C left). This may be because slower bump movement decreases the firing irregularity of flank neurons, which deviates from the Poisson firing assumption of our theory (see also Discussion). However, given the simplifying assumptions needed to derive the theory, the match to the spiking network is surprisingly accurate.

Short-term plasticity controls drift

Having established that our theory is able to predict the effect of STP on diffusion, as well as drift for a single instantiation of random connectivity, we wondered how different sources of heterogeneity (frozen noise) would influence the drift of the bump. We considered two sources of heterogeneity: First, random connectivity as introduced above, and second, heterogeneity of the leak reversal potential parameters of excitatory neurons: leak reversal potentials of excitatory neurons are given by $V_L + \Delta_L$, where Δ_L is normally distributed with zero mean and standard deviation σ_L [36]. The resulting fields can be calculated by calculating the resulting perturbations to the firing rates of neurons by Eq. (8) (see *Frozen noise* in Materials and Methods for details).

The theory developed so far allowed us to predict drift-fields for a given realization of frozen noise, controlled by the noise parameters p (for random connectivity) and σ_L (for heterogeneous leak reversal-potentials) (see S3 Fig for a comparison of predicted drift fields to those measured in simulations for varying STP parameters and varying strengths of frozen noises). We wondered, whether we could take the level of abstraction of our theory one step further, by predicting the magnitude of drift fields from the frozen noise parameters only, independently of a specific realization. First, the expectation of drift fields under the distributions of the frozen noises vanishes for any given position: $\langle A(\varphi)\rangle_{\text{frozen}} = 0$, where the expectation $\langle\,.\,\rangle_{\text{frozen}}$ is taken over both noise parameters. We thus turned to the expected squared magnitude of drift fields under the distributions of these parameters (see *Squared field magnitude* in Materials and Methods for the derivation):

$$\left\langle A^2\right\rangle_{\text{frozen}} = \frac{1}{S^2}\sum_i C_i^2 \left(\frac{\left(\phi'_{0,i}\right)^2}{N_E^2}\left(\frac{1}{p}-1\right)\sum_j \left(s_{0,j}\right)^2\left(w_{ij}^{\text{EE}}\right)^2 + \left(\frac{d\phi_{0,i}}{d\Delta_i^L}\right)^2\sigma_L^2 \right), \quad (9)$$

where $s_{0,j}$ is the steady-state synaptic activation. Here, we introduced the derivatives of the input-output relation with respect to the noise sources that appear in Eq. (8): $\phi'_{0,i} = \frac{dF}{dJ}(J_{0,i}(\varphi))$ is the derivative with respect to the steady state synaptic input, and

$\frac{d\phi_{0,i}}{d\Delta_i^L}$ is the derivative with respect to the perturbation in the leak potential. In *Squared field magnitude* in Materials and Methods, we show that Eq. (9) is independent of the center position φ, and can be estimated from simulations as the variance of the drift field across positions, averaged over an ensemble of network instantiations.

We defined the root of the expected squared magnitude of Eq. (9) as the *expected field magnitude*:

$$\sqrt{\langle A^2 \rangle_{\text{frozen}}}. \tag{10}$$

This quantity predicts the magnitude of the deviations of drift-fields from zero that are expected from the parameters that control the frozen noise – in analogy to the standard deviation for random variables, it predicts the standard deviation of the fields. To compare this quantity to simulations, we varied both heterogeneity parameters. First, the connectivity parameter p was varied between 0.25 and 1. Second, for heterogeneities in leak reversal-potentials, we chose values for the standard deviation σ_L of leak-reversal potentials between $0mV$ and $1.5mV$, which lead to a similar range of drift magnitudes as those of randomly connected networks. For each combination of heterogeneities and STP parameters (networks had either random connections or heterogeneous leaks) we then realized $18 - 20$ networks, for which we simulated 400 repetitions of $6.5s$ of delay activity each (20 uniformly spaced positions of the initial cue). We then estimated the drift-field numerically by recording displacements of bump centers along their trajectories (as in Fig. 2A, B) and measured the standard deviation of the resulting fields across all positions.

Similar to the analysis of diffusion above, we find that facilitation and depression elicit antagonistic control over the magnitude of drift fields. In both simulations and theory, we find (Fig. 4A,B) that the expected field magnitude *decreases* as the effect of facilitation is *increased* from unfacilitated networks ($U = 1$) through intermediate levels of facilitation ($U = 0.4$) to strongly facilitating networks ($U = 0.1$). Our theory predicts this effect surprisingly well, which we validated twofold (as for the diffusion magnitude). First, we used Eq. (10) with all parameters and coefficients estimated from each spiking simulation separately (Fig. 4A,B, plus-signs and crosses). Second, we extrapolated the theoretical prediction by using coefficients in Eq. (9) measured from the unfacilitated reference network only ($U = 1, \tau_x = 150ms$) but changed the facilitation and heterogeneity parameters (Fig. 4A,B, dashed lines). The largest differences between the extrapolated and full theory are seen for $U < 1$ and randomly connected networks ($p < 1$), which we found to result from the fact that bump shapes for these networks tended to be slightly reduced under random and sparse connectivity (e.g. the top firing rate is reduced to $\sim 35Hz$ for $U = 0.1, p = 0.25$). Generally, as noise levels increase, our theory tends to over-estimate the squared magnitude of fields, since we rely on a linear expansion of perturbations to the firing rates to calculate fields (Eq. (8)). Such deviations are expected as the magnitude of firing rate perturbations increases, and could be counter-acted by including higher-order terms. Since in the theory facilitation (and depression) only scales the firing rate perturbations (Eq. (7)), these deviations can also be observed across facilitation parameters. Finally, we performed a similar analysis to investigate the effect of short-term depression on drift fields. Here, we varied the depression time constant τ_x for randomly connected networks with $p = 0.6$, by simulating networks with combinations of short-term plasticity parameters from $U \in \{0.1, 0.4, 0.8\}$ and $\tau_x \in \{120ms, 160ms, 200ms\}$ (Fig. 4C). We find that an increase of the depression time constant leads to increased magnitude of drift fields, which again is well predicted by our theory.

Fig 4. Drift field magnitude is controlled by short-term plasticity. A Expected magnitude of drift fields as a function of the sparsity parameter p of recurrent excitatory-to-excitatory connections. Dots are the standard deviation of fields estimated from 400 trajectories (see main text) of each network, averaged over $18-20$ realizations for each noise parameter and facilitation setting (error bars show 95% confidence of the mean). Theoretical predictions (dashed lines) are given by Eq. (10) extrapolated from the reference network ($U = 1, \tau_x = 150$). For validation, we also estimated Eq. (10) with coefficients measured from each simulated network separately (plus signs). The depression time constant was $\tau_x = 150$ms. **B** Same as in panel A, with heterogeneous leak-reversal potentials as the source of frozen noise. Validation predictions are plotted as crosses. **C** Same as in panels A,B but varying the depression time constant τ_x for a fixed level of frozen noise (random connectivity, $p = 0.6$). In all panels, the facilitation time constant was $\tau_u = 650$ms.

Short-term plasticity controls memory retention

The theory developed in previous sections shows that diffusion and drift of the bump center φ are controlled antagonistically by short-term depression and facilitation. In a working memory setup, we can view the attractor dynamics as a noisy communication channel [54] that maps a set of initial positions $\varphi(t = 0s)$ (time of the cue offset in the attractor network) to associated final positions $\varphi(t = 6.5s)$, after a memory retention delay of $6.5s$. We used the distributions of initial and (associated) final positions to investigate the combined impact of diffusion and drift on the retention of memories (Fig. 5A). Because of diffusion, distributions of positions will widen over time, which degrades the ability to distinguish different initial positions of the bump center (Fig. 5A, top). Additionally, directed drift of the dynamics will contract distributions of different initial positions around the same fixed points, making them essentially indistinguishable when read out (Fig. 5A, bottom).

As a numerical measure of this ability of such systems to retain memories over the delay period, we turned to mutual information (MI), which provides a measure of the amount of information contained in the readout position about the initially encoded position [55, 56]. To measure MI from simulations (see *Mutual information measure* in Materials and Methods), we analyzed network simulations for varying short-term facilitation parameters (U) and magnitudes of frozen noises (p and σ_L) (same data set as Fig. 4A,B). We recorded the center positions encoded in the network at the time of cue-offset ($t = 0$) and after $6.5s$ of delay activity, and used binned histograms (100 bins) to calculate discrete probability distributions of initial ($t = 0$) and final positions ($t = 6.5$). For each trajectory simulated in spiking networks, we then generated a trajectory starting at the same initial position by using the Langevin equation Eq. (4) that describes the drift and diffusion dynamics of center positions. The MI calculated from the resulting distributions of final positions (again at $t = 6.5$) for each network serve as the theoretical prediction for each network. As a reference, we used the spiking network without facilitation ($U = 1, \tau_u = 650ms, \tau_x = 150ms$) and no frozen noises ($p = 1$, $\sigma_L = 0mV$) and normalized the MI of all other networks (both for spiking simulations and theoretical predictions) with respect to the reference, yielding the measure of *relative MI* presented in Fig. 5B-E.

We found that the relative MI decreased compared to the reference network as network heterogeneities were introduced (Fig. 5B, green). This was expected, since directed drift caused by heterogeneities leads to a loss of information about initial positions. There were two effects of increased short-term facilitation (by decreasing the parameter U). First, diffusion was reduced, which was visible in a vertical shift of the relative MI for facilitated networks (Fig. 5A, orange and blue, at 0 heterogeneity).

Fig 5. Short-term facilitation increases memory retention. A Illustration of the effects of diffusion (top) and additional drift (bottom) on the temporal evolution of distributions of initial positions $p(\text{start})$ towards distributions of final positions $p(\text{end})$ over $6.5s$ of delay activity. The bump is always represented by its center position φ. Two peaks in the distribution of initial positions $\varphi(0)$ and their corresponding final positions $\varphi(6.5)$ are highlighted by colors (purple, red), together with example trajectories of the center positions. Top: Diffusion symmetrically widens the initial distribution. Bottom: Strong drift towards one single fixed point of bump centers ($\varphi = 0$) makes the origin of trajectories indistinguishable. **B** Normalized mutual information (MI, see text for details) of distributions of initial and final bump center positions in working memory networks for different STP parameters and heterogeneity parameters(blue: strong facilitation, see legend in panel D). Dots and triangles are average MI ($18 - 20$ realizations, error bars show 95% CI) obtained from spiking network simulations. Lines show average MI calculated from Langevin dynamics for the same networks, repetitions and realizations (see text, shaded area shows 95% CI). Heterogeneity parameters are σ_L (triangles, in units of mV) and $1 - p$ (circles), where p is the connection probability. **C** Expected displacement $|\Delta\varphi|\,(1s)$ for the same networks as in panel B. Dashed lines indicate displacement induced by diffusion only ($\sqrt{B}$), solid lines show the total displacement (including displacement due to drift, calculated as the expected field magnitude $\sqrt{\langle A^2\rangle_{\text{frozen}}}$). **D** Same as panel B, with x-axis showing the expected field magnitude. **E** Same as panel B, with x-axis showing the expected displacement. In panels B-D, all facilitation parameters except U were kept constant at $\tau_u = 650\text{ms}, \tau_x = 150\text{ms}$

Second, the effects of frozen noise decreased with increasing facilitation, which was visible in the slopes of the MI decrease (see also S4 Fig). The MI obtained by integration of the Langevin equations (see above) matched those of the simulations well (Fig. 5A, lines). From earlier results, we expected the drift-fields to be slightly over-estimated by the theory as the heterogeneity parameters increase (Fig. 4), which would lead to an under-estimation of MI. We did observe this here, although for $U = 1$ the effect was slightly counter-balanced by the under-estimated level of diffusion (cf. Fig. 3A, right), which we expected to increase the MI. For networks with stronger facilitation ($U = 0.1$), we systematically over-estimated diffusion (cf. Fig. 3, left), and therefore under-estimated MI.

Using our theory, we were able to simplify the functional dependence between MI, short-term plasticity, and frozen noise. Combining the effects of both diffusion and drift into a single quantity for each network, we replaced the field $A(\varphi)$ by our theoretical prediction $\sqrt{\langle A^2\rangle_{\text{frozen}}}$ in Eq. (4) and forward integrated the differential equation for a time interval $\Delta t = 1s$, to arrive at the *expected displacement* in $1s$:

$$|\Delta\varphi|\,(1s) = \sqrt{\langle A^2\rangle_{\text{frozen}}} \cdot 1s + \sqrt{B \cdot 1s}. \tag{11}$$

This quantity describes the expected absolute value of displacement of center positions during $1s$: it increases as a function of the frozen noise distribution parameters (Fig. 5C), but even in the absence of frozen noise it is nonzero due to diffusion. Plotting the MI data in dependence of the first term only ($\sqrt{\langle A^2\rangle_{\text{frozen}}}$), shows that the MI curves collapse onto a single curve for each facilitation parameter (Fig. 5D). Finally, plotting the MI data against $|\Delta\varphi|\,(1s)$ we find that all data collapse on to nearly a single curve (Fig. 5E). Thus, the effects of the two sources of frozen noise (corresponding to $\langle A^2\rangle_{\text{frozen}}$) and diffusion (corresponding to B) are unified into a single quantity $|\Delta\varphi|\,(1s)$.

We performed the same analyses on a large set of network simulations with fixed random connectivity ($p = 0.6$) and varying STP parameters for both depression (τ_x)

Fig 6. Short-term depression decreases memory retention. A Same as Fig. 5B, for network simulations with varying τ_x and U (see legend in panel B). MI is normalized to the same value as there. **B** Same as panel A, with x-axis showing the expected displacement. Light gray data points and lines are the data plotted in Fig. 5E. The facilitation time constant was kept constant at $\tau_u = 650$ms.

and facilitation (U) (same data set as in Fig. 4C). Increasing the short-term depression time constant τ_x leads to decreased relative MI with a positive offset induced through stronger facilitation (Fig. 6A, blue line). Calculating the expected displacement for these network configurations collapsed the data points mostly onto the same curve as earlier (Fig. 6B). For strong depression combined with weak facilitation ($\tau_x = 200ms$, $U = 0.8$), the drop-off of the relative MI saturates earlier, indicating that for these strongly diffusive networks the effect on MI may not be sufficiently captured by its relationship to $|\Delta\varphi|\,(1s)$.

Linking theory to experiments: distractors & network size

The large level of abstraction of our theory condenses the complex dynamics of spiking bump attractor networks into a high-level description of a few macroscopic features, which in turn allows matching the theory to behavioral experiments. Here, we demonstrate how such quantitative links could be established using two different features: 1) the sensitivity of the working memory circuit to distractors, and 2) the stability of working memory expressed by the expected displacement. We stress that our model is a simplified description of biological circuits (see Discussion), and thus, at the current level of abstraction, should be seen as a proof of principle rather than a means of deriving accurate quantitative predictions.

Predicting the sensitivity to distractor inputs

In a biological setting, drifts introduced by network heterogeneities (frozen noise) could be significantly reduced by (long-term) plasticity [36]. To measure the intrinsic stability of continuous attractor models, earlier studies [11, 46, 57] have alternatively proposed to use *distractor inputs* (Fig. 7A): providing a short external input centered around a position φ_D to the network, the center position of an existing bump state will be biased towards the distracting input, with stronger biases appearing for closer distractors. Distractors forcibly expose the working memory system's intrinsic time scale by creating externally induced drifts, thereby testing its sensitivity (in terms of distractor distance), and they are equally implementable in behavioral experiments [57].

Our theory can readily yield quantitative predictions for the distractor paradigm. To accommodate distractor inputs in the theory, we assume that they cause some units i to fire at elevated rates $\phi_{0,i} + \Delta\phi_i$, which will introduce a drift field according to Eq. (7) (Fig. 7A, purple dashed line). The resulting dynamics (Eq. (4)) of diffusion and drift during the presentation of the distractor input then allow us to calculate the expected shift of center positions as a function of all network parameters, including those of short-term plasticity. Repeating this paradigm for varying positions of the distractor inputs (see *Distractor analysis* in Materials and Methods for details), our theory predicts that strong facilitation will strongly decrease both the effect and radial reach of distractor inputs (Fig. 7B, blue), when compared to the unfacilitated system (Fig. 7B, green) – in qualitative agreement with simulation results involving a related (cell-intrinsic) stabilization mechanism [46]. Conversely, we predict that longer recovery from short-term depression tends to increase the sensitivity to distractors (Fig. 7C). The displacement caused by each distractor input is given integrating the resulting dynamics

Fig 7. Effect of short-term plasticity on distractor inputs. A While a bump ("Bump") is centered at an initial angle φ_0 (chosen to be 0), additional external input causes neurons centered around the position φ_D to fire at elevated rates ("Distractor"). The theory predicts the shape and magnitude of the induced drift field ("Field") and the mean bump center φ_1 after 250ms of distractor input. Gray trajectories are example simulations of bump centers of the corresponding Langevin equation Eq. (4). **B** Mean final positions φ_1 of bump centers (1000 repetitions, shaded areas show 1 standard deviation) as a function of the distractor input location φ_D. Increased short-term facilitation (blue: strong facilitation, $U = 0.1$; orange: intermediate facilitation, $U = 0.4$; green: no facilitation $U = 1$) leads to less displacement due to the distractor input. Other STP parameters were kept constant at $\tau_u = 650\text{ms}, \tau_x = 150\text{ms}$. **C** Same as panel B, for three different depression time constants τ_x, while keeping $U = 0.8, \tau_u = 650\text{ms}$ fixed. **D** Same as panel B, with a broader bump half-width ($\sigma_g = 0.8\text{rad} \approx 45.8\,\text{deg}$). All other panels use the same bump half-width as in the rest of the study ($\sigma_g = 0.5\text{rad} \approx 28.7\,\text{deg}$) (see S1 Fig).

of Eq. (4) over the stimulus duration. As such, the magnitude of the displacement will increase both with the amplitude and the duration of the distractor input. Finally, our theory demonstrates that the bump shape, importantly bumps with a broader width, can also significantly increase the effect and radial reach of distractor inputs (Fig. 7D).

Constraining the size of working memory networks

The simple theoretical measure of expected displacement $|\Delta\varphi|\,(1s)$ introduced in the last section can be related to behavioral experiments: a value of $|\Delta\varphi|\,(1s) = 1.0\,\text{deg}$ lies in the upper range of experimentally reported deviations due to diffusive and systematic errors in behavioral studies [58, 59]. What are the microscopic circuit compositions that can attain such a (high) level of working memory stability? In particular, since an increase in network size can reduce diffusion [11] and the effects of random heterogeneities [16, 36, 38, 45], we turned to the question: *which networks sizes would be needed to yield this level of stability in a one-dimensional continuous memory system*?

To address the question of network size, we extended our theory to include the size N of the excitatory population as an explicit parameter (see *System size scaling* in Materials and Methods for details). We find (see Eq. (59)) that the expected field magnitude $\sqrt{\langle A^2\rangle_{\text{frozen}}}$ (cf. Eq. (10)) scales with the connectivity parameter p and the system size N to leading order as $1/(\sqrt{p}N)$, whereas the perturbations due to heterogeneity of leak potentials scale as $1/\sqrt{N}$, both in accordance with earlier results [16, 36, 38, 45]. For the diffusion scale $\sqrt{B}$ (cf. Eq. (5)) we find a scaling as as $1/\sqrt{N}$, also in agreement with earlier work [11, 14, 36, 39, 45]. Using numerical coefficients in Eq. (4) extracted from the spiking simulation of a reference network ($U = 1, \tau_x = 150$ and $N_E = 800$), we extrapolated the theory by changing the system size N and short-term plasticity parameters. We then constrained parameters of our theory by published data (Table 1). Short-term plasticity parameters were based on two groups of strongly facilitating synapses found in a study of mammalian (ferret) prefrontal cortex [60]. The same study reported a general probability $p = 0.12$ of pyramidal cells to be connected. However, for pairs of pyramidal cells that were connected by facilitating synapses, the study found a high probability of reciprocal connections ($p_{rec} = 0.44$): thus if neuron A was connected to neuron B (with probability p), neuron B was connected to neuron A with high probability (p_{rec}), resulting in a non-random connectivity. To approximate this in the random connectivities supported by our theory, we evaluated a second, slightly elevated, level of random connectivity, that has the same mean connection probability as the non-random connectivity with these additional

reciprocal connections: $p + p \cdot p_{rec} = 0.1728$. For the standard deviation of leak reversal-potentials σ_L, we used values measured in two studies [61, 62].

Table 1. Upper bounds on system-sizes for stable continuous attractor memory in prefrontal cortex.

STP parameters	$\Delta\varphi(1s)$	p	σ_L	**Network size** N
$U = 0.17$	1.0 deg	0.12	1.7mV [61, RS]	79 504
$\tau_u = 563ms$	[58, 59]	[60]	2.4mV [62, fa-RS]	127 465
$\tau_x = 242ms$		0.1728	1.7mV	79 047
[60, E1b]		[60]	2.4mV	127 205
	0.5 deg [58]	0.12	2.4mV	507 607
$U = 0.35$	1.0 deg	0.12	1.7mV	102 292
$\tau_u = 482ms$			2.4mV	163 896
$\tau_x = 163ms$		0.1728	1.7mV	101 836
[60, E1a]			2.4mV	163 638
	0.5 deg	0.12	2.4mV	653 350

Theoretical predictions of Eq. (11) optimized for the number of excitatory neurons N that are needed to achieve a given level of expected displacement $|\Delta\varphi|\,(1s)$ under given parameters of short-term plasticity and frozen noises. RS: regular spiking pyramidal cells, fa-RS: fast-adapting regular spiking pyramidal cells.

The resulting theory makes quantitative predictions for combinations of network size N and all other parameters that yield the desired levels of working memory stability (Table 1, see also S5 Fig). Network sizes were all smaller than 10^6 neurons, with values depending most strongly on the value of the facilitation parameter U and the magnitude of the leak reversal-potential heterogeneities σ_L. Since the expected field magnitude scales weakly ($1/\sqrt{p}$) with the recurrent connectivity p, increasing p lead only to comparatively small decreases in the predicted network sizes. Finally, we see that the increasing the reliability of networks comes at a high cost: decreasing the expected displacement to $|\Delta\varphi|\,(1s) = 0.5\,\text{deg}$ [58] increases the required number of neurons by nearly a number of 4 for both facilitation settings we investigated. Nevertheless, these network sizes still lie within anatomically reasonable ranges [63]. In summary, our theory predicts that, given the simplifying assumptions of our models, continuous attractor networks with realistic values for the strength of facilitation and depression of recurrent connections could achieve sufficient stability, even in the presence of a realistic degree biological variability.

Discussion

We presented a theory of drift and diffusion in continuous working memory models, exemplified on a one-dimensional ring attractor model. Our framework extends earlier approaches calculating the effects of noise by projection onto the attractor manifold [37, 39, 40] by including the effects of short-term plasticity. Our approach further generalizes earlier work on drift in continuous attractors with short-term plasticity [38] to include diffusion and the dynamics of short-term depression. The theory predicts that facilitation and depression play opposite roles in making continuous attractors robust against the influences of both dynamic noise (introduced by spiking variability) and frozen noise (introduced by biological variability). We have confirmed the quantitative predictions of our theory in simulations of a ring-attractor implemented in a spiking network model with synaptic facilitation and depression, and found theory and simulation to be quantitatively in good agreement.

In *Short-term plasticity controls memory retention*, we demonstrated the effects that STP has through drift and diffusion on the information retained in continuous working memory. Using our theoretical predictions of drift and diffusion we were able to derive the expected displacement $|\Delta\varphi|$ as a function of STP parameters and the frozen noise parameters, which provides a simple link between the resulting Langevin dynamics of bump centers and mutual information (MI) as a measure of working memory retention. Our results can be generalized in several directions. First, the choice of $1s$ of forward integrated time for $|\Delta\varphi|$ (Eq. (11)) was arbitrary. While a choice of $\sim 2s$ lets the curves in Fig. 5E collapse slightly better, we chose $1s$ to avoid further heuristics. Second, we expect values of MI to decrease as the length of the delay period is increased. Our choice of $6.5s$ is comparable to delay periods often considered in behavioral experiments (usually 3-6s) [59, 64, 65]. However, a more rigorous link between the MI measure and the underlying attractor dynamics would be desirable. Indeed, for noisy channels governed by Fokker-Planck equations, this might be feasible [66], but goes beyond the scope of this work.

In *Linking theory to experiments: distractors & network size*, we demonstrated that the high-level description of the microscopic dynamics obtained by our theory allows its parameters to be constrained by experiments. Considering that our model is a simplified description of its biological counterparts (see next paragraph), these demonstrations are to be seen as a proof of principle, which will not be able to yield biologically accurate predictions in its current level of abstraction. However, since distractor inputs can be implemented in silico as well as in behavioral experiments (see e.g. [57]), they could eventually provide a quantitative link between the continuous attractor models and working memory systems, by matching the resulting distraction curves. Given the simplifying assumptions, our theory provides a direct map between underlying network parameters and their combined effects on distraction sensitivity, facilitating previous approaches in which these distraction curves had to be extracted through repeated microscopic simulations for single parameter settings [46]. We further used our theory to derive bounds on network parameters, in particular the size of networks, that lead to "tolerable" levels of drift and diffusion in the simplified model. For large magnitudes of frozen noise our theory tends to over-estimate the expected magnitude of drift-fields slightly (cf. Fig. 4). Thus, we expect the predictions made here to be upper bounds on network parameters needed to achieve a certain expected displacement. Finally, while the predictions of our theory might deviate from biological networks, they could be applied to accurately characterize the stability of, and the effects of inputs to, bump attractor networks implemented in neuromorphic hardware for robotics applications [67].

Our results show, to our knowledge for the first time, that strong facilitation (small values of U) does not only slow down directed drift [38], but also efficiently suppresses diffusion in continuous attractor models. However, in delayed response tasks involving saccades, that presumably involve continuous attractors in the prefrontal cortex [11, 22], one does observe an increase of variability in time [64]: both quickly accumulating systematic errors (alike drift) [59] and more slowly increasing variable errors (with variability growing linear in time, alike diffusion) [58] appear. Indeed, there are several other possible sources of variability in cortical working memory circuits, which we did not consider here. For example, noisy synaptic transmission and STP [68] or heterogeneous STP parameters as found in [60]. These will probably induce further drift and diffusion and the effectiveness of facilitation as a stabilizing mechanism in coping with them remains to be investigated. For simplicity, we also excluded AMPA currents from the recurrent excitatory interactions, which are found in biological neuronal networks [11]. However, since STP acts by presynaptic scaling of neurotransmitter release, it will act symmetrically on both AMPA and NMDA receptors, which we thus expect to lead to similar dynamical effects of STP as those reported here and which

might be amenable to a similar analytical approach. Additionally, variable errors might be introduced elsewhere in the pathway between visual input and motor output (but see [69]) or by input from of other noisy local circuits during the delay period [70].

Several additional dynamical mechanisms might also influence the stability of continuous attractor working memory circuits. For example, intrinsic neuronal currents that modulate the neuronal excitability [46] or firing-rate adaptation [71] have been shown to affect bump stability. It should be noted that many such dynamical processes could be accommodated in our theoretical approach, by including their linearized dynamics in the calculation of the projection vector (cf. *Projection of dynamics onto the attractor manifold* in Materials and Methods). Fast corrective inhibitory feedback has also been shown to stabilize spatial working memory systems (in balanced networks) [72]. It has been shown that, on the timescale of hours to days, homeostatic processes can be used to counteract the drift introduced by frozen noise [36]. Finally, inhibitory connections that are distance-dependent [11] and show short-term plasticity [73] could also influence bump dynamics.

We have focused here on ring-attractor models that obtain their stable firing-rate profile due to perfectly symmetric connectivity. Our approach can also be employed to analyze ring-attractor networks with short-term plasticity, in which weights show (deterministic or stochastic) deviations from symmetry (see *Frozen noise* in Materials and Methods for stochastic deviations). Although not investigated here, continuous line-attractors arising through a different weight-symmetry should be amenable to similar analyses [39]. Finally, it should be noted that adequate structuring of the recurrent connectivity can also positively affect the stability of continuous attractors [14]. For example, translational asymmetries included in the structured heterogeneity can break the continuous attractor into several isolated fixed points, which can lead to decreased diffusion along the attractor [56].

We provided evidence that short-term synaptic plasticity controls the sensitivity of attractor networks to both fast diffusive and frozen noise. Control of short-term plasticity via neuromodulation [74] would thus represent an efficient "crank" for adapting the timescale of computations of such networks. By changing the properties of presynaptic calcium entry [75], inhibitory modulation mediated via $\mathrm{GABA_B}$ and adenosine $\mathrm{A_1}$ receptors can lead to increased facilitatory components in rodent cerebellar [76] and avian auditory synapses [77]. Dopamine, serotonin and noradrenaline have all been shown to differentially modulate short-term depression (and facilitation when blocking GABA receptors) at sensorimotor synapses [78]. Interestingly, next to short-term facilitation on the timescale of seconds, other dynamic processes up-regulate recurrent excitatory synaptic connections in prefrontal cortex [60]: synaptic augmentation and post-tetanic potentiation operate on longer time scales (up to tens of seconds), and might be able to support working memory function [79]. While the long time scales of these processes might again render putative short-term memory networks inflexible, there is evidence that they might also be under tight neuromodulatory control [80]. Neuromodulation could also generally provide higher flexibility (see above) to a facilitation-stabilized working memory system, by allowing for more dynamical control of the rigidity of memory representations [46].

Comparison to earlier work

Similar to an earlier theoretical approach using a simplified rate model [38], we find that the slowing of drift by facilitation depends mainly on the facilitation parameter U, while the time constant τ_u has a less pronounced effect. While the approach of [38] relied on the projection of frozen noise onto the derivative of the first spatial Fourier mode of the bump shape along the ring, here we reproduce and extend this result (1) for arbitrary neuronal input-output relations and (2) a more detailed spatial projection that involves

the synaptic dynamics and the bump shape. While, our theory can also accommodate noisy recurrent connection weights as frozen noise, as used in in [38] (see *Frozen noise* in Materials and Methods for derivations), the drifts generated by these heterogeneities were generally small compared to diffusion and the other sources of heterogeneity.

A second study investigated short-term facilitation and showed that it reduces drift and diffusion in a spiking network, for a fixed setting of U (although the model of short-term facilitation differs slightly from the one employed here) [46]. Contrary to what we find here, these authors find that an increase in τ_u leads to increased diffusion, while we find that an increase over the range they investigated ($\sim 0.5s - 4s$) would decrease the diffusion by a factor of nearly two. More precisely, for our shape of the bump state (which we keep fixed) we predict a reduction from ~ 26 to ~ 16 deg^2/s for a similar setting of facilitation U. These differences might arise from an increasing width of the bump attractor profile for growing facilitation time constants in [46], which would then lead to increased diffusion in our model. Whether this effect persists under the two-equation model of saturating NMDA synapses used there remains to be investigated. Finally, increasing the time constant of recurrent NMDA conductances has been shown to also reduce diffusion [46], in agreement with our theory, according to which the normalization constant S increases with τ_s [39].

In a study that investigated only a single parameter value for depression ($\tau_x = 160ms$, no facilitation) in a spiking network similar to the one investigated here [44], the authors observed no apparent effect of short-term depression on the stability of the bump. In contrast, we find that stronger short-term depression will indeed increase both diffusion and directed drift along the attractor. Our result agrees qualitatively with earlier studies in rate models, which showed that synaptic depression, similar to neuronal adaptation [10, 81], can induce movement of bump attractors [42, 43, 82, 83]. In particular, the study of [42] showed that for simpler rate models a regime exists where the bump state moves with constant speed along the attractor manifold. We did not find any such directed movement in our networks, which could be due to fast spiking noise stabilizing the directed movement of bumps observed in noise-free rate models, as shown in [81]: there, additive fast noise was able to cancel directed bump movement caused by single neuron adaptation.

Extensions & Shortcomings

The coefficients of Eq. (4) give clear predictions as to how drift and diffusion will depend on the shape of the bump state and the neural transfer function F. The relation is not trivial, since the pre-factors C_i and the normalization constant S also depend on the bump shape. For the diffusion strength Eq. (5), we explored this relation numerically, by artificially varying the shape of the firing rate profile (while extrapolating other quantities). Although a more thorough analysis remains to be performed, a preliminary analysis shows (see S6 Fig) that diffusion increases both with bump width and top firing rate, consistent with earlier findings [11, 32].

In *Linking theory to experiments: distractors & network size*, we have demonstrated that our theory can be used to predict the shape and effect of drift fields that are generated by localized external inputs due to distractor inputs. More generally, any localized external input (excitatory or inhibitory) will cause a deviation $\Delta\phi_i$ from the steady-state firing rates, which, in turn, generates a drift field by Eq. (7). This could predict the strength and location of external inputs that are needed to induce continuous shifts of the bump center at given speeds, for example when these attractor networks are designed to track external inputs (see e.g. [10, 84]). It should be noted that in our simple approximation of this distractor scheme, we assume the system to remain at approximately steady-state, i.e. that the bump shape is unaffected by the additional external input, except for a shift of the center position. For example, we expect

additional feedback inhibition (through the increased firing of excitatory neurons caused by the distractor input) to decrease bump firing rates. A more in depth study and comparison to simulations will be left for further work.

The spiking networks we analyzed here are tuned to display balanced inhibition and excitation in the inhibition dominated uniform state [85, 86], while the bump state relies on positive currents, mediated through strong recurrent excitatory connections (cf. [44] for an analysis). Similar to other spiking network models of this class, this mean–driven bump state shows relatively low variability of neuronal inter-spike-intervals, because near the bump center the mean input is close to (or even above) the firing threshold [87, 88] (see also next paragraph). While the decreased variability appears for neurons in the center of the firing rate profile, we found that neurons at its flanks still display variable firing, with statistics close to that expected of spike trains with Poisson statistics (see S7 Fig), which may be because the flank's position slightly jitters. Since the non-zero contributions to the diffusion strength are constrained to these flanks (cf. Fig. 1D), the simple theoretical assumption of Poisson statistics of neuronal firing still matches the spiking network quite well. As discussed in *Short-term plasticity controls diffusion*, we find that our theory over-estimates the diffusion as bump movement slows down for small values of U – this may be due to a decrease in firing irregularity in stable bumps, at which the Poisson assumption becomes inaccurate.

More recent bump attractor approaches allow networks to perform working memory function with a high firing variability also during the delay period [3], in better agreement with experimental evidence [89]. These networks show bi-stability, where both stable states show balanced excitation and inhibition [87] and the higher self-sustained activity in the delay activity is evoked by an increase in fluctuations of the input currents (noise-driven) rather than an increase in the mean input [90]. This was also reported for a ring-attractor network (with distance-dependent connections between all populations), where facilitation and depression are crucial for irregularity of neuronal activity in the self-sustained state [45]. Application of our approach to these setups is left for future work.

Materials and methods

Analysis of drift & diffusion with STP

For the following, we define a concatenated $3 \cdot N$ dimensional column vector of state variables $\mathbf{y} = \left(\mathbf{s}^T, \mathbf{u}^T, \mathbf{x}^T\right)^T$ of the system Eq. (3). Given a (numerical) solution of the stable firing rate profile $\vec{\phi}_0$ we can calculate the stable fixed point of this system by setting the l.h.s. of Eq. (3) to zero. This yields steady-state solutions for the synaptic activations, facilitation and depression variables $\mathbf{y}_0 = (\mathbf{s}_0, \mathbf{u}_0, \mathbf{x}_0)$:

$$\begin{aligned} s_{0,i} &= \tau_s u_{0,i} x_{0,i} \phi_{0,i}, \\ u_{0,i} &= U \frac{1 + \tau_u \phi_{0,i}}{1 + U \tau_u \phi_{0,i}}, \\ x_{0,i} &= \frac{1 + U \tau_u \phi_{0,i}}{1 + U \left(\tau_u \phi_{0,i} + \tau_u \tau_x \phi_{0,i}^2 + \tau_x \phi_{0,i}\right)}. \end{aligned} \tag{12}$$

We then linearize the system Eq. (3) at the fixed point $\mathbf{y}_0$, introducing a change of variables consisting of perturbations around the fixed point: $\mathbf{y} = \mathbf{y}_0 + \delta\mathbf{y} = \mathbf{y}_0 + \left(\delta\mathbf{s}^T, \delta\mathbf{u}^T, \delta\mathbf{x}^T\right)$ and $\phi_i = \phi_{0,i} + \delta\phi_i$. To reach a self-consistent linear system, we further assume a separation of time scales between the neuronal dynamics and the synaptic variables, in that the neuronal firing rate changes as an immediate function of the (slow) input. This allows replacing

$\delta\phi_i = \frac{d\phi_i}{dJ_i}\Big|_{J_{0,i}} \sum_j \frac{dJ_i}{ds_j}\delta s_j = \phi'_{0,i}\sum_j w_{ij}\delta s_j$, where we introduce the shorthand $\phi'_{0,i} \equiv \frac{d\phi_i}{dJ_i}\Big|_{J_{0,i}}$. Finally, keeping only linear orders in all perturbations, we arrive at the linearized system equivalent of Eq. (3):

$$\dot{\delta\mathbf{y}} = \begin{pmatrix} -\frac{1}{\tau_s}\mathbb{I} + D\left(\mathbf{u}_0 \cdot \mathbf{x}_0 \cdot \vec{\phi}'_0\right)W & D\left(\vec{\phi}_0 \cdot \mathbf{x}_0\right) & D\left(\vec{\phi}_0 \cdot \mathbf{u}_0\right) \\ UD\left((1-\mathbf{u}_0)\cdot\vec{\phi}'_0\right)W & -\frac{1}{\tau_u}\mathbb{I} - UD\left(\vec{\phi}_0\right) & 0 \\ -D\left(\mathbf{u}_0 \cdot \mathbf{x}_0 \cdot \vec{\phi}'_0\right)W & -D(\mathbf{x}_0 \cdot \vec{\phi}_0) & -\frac{1}{\tau_x}\mathbb{I} - D\left(\vec{\phi}_0 \cdot \mathbf{u}_0\right) \end{pmatrix}\delta\mathbf{y}$$
$$\equiv K\delta\mathbf{y} \tag{13}$$

Here, dots between vectors indicate element-wise multiplication, the operator $D : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ creates diagonal matrices from vectors, and $W = (w_{ij})$ is the synaptic weight matrix of the network.

Projection of dynamics onto the attractor manifold

To project the dynamical system Eq. (13) onto movement of the center position φ of the firing rate profile, we assume that N is large enough to treat the center position φ as a continuous variable. We also assume that the network implements a ring-attractor: the system dynamics are such that the firing rate profile $\vec{\phi}_0$ can be freely shifted to different positions along the ring, changing the center position φ, while retaining the same shape. All other possible directions of change in this system are assumed to be constrained by the system dynamics. In the system at hand, this implies that the matrix K of Eq. (13), which captures the linearized dynamics around any of these fixed points, will have a *zero eigenvalue* corresponding to the eigenvector of a change of the dynamical variables under a change of position φ, while all other eigenvalues are negative [39].

Formally, the column eigenvector to the eigenvalue 0 is given by changes in the state variables as the bump center position φ is translated along the manifold:

$$e_r = \frac{d\mathbf{y}_0}{d\varphi} = \left(\frac{d\mathbf{s}_0}{d\varphi}^T, \frac{d\mathbf{u}_0}{d\varphi}^T, \frac{d\mathbf{x}_0}{d\varphi}^T\right)^T. \tag{14}$$

Let e_l be the associated row left-eigenvector (also to eigenvalue 0) of K, normalized such that:

$$e_l \cdot e_r = 1. \tag{15}$$

In Section 1 of S1 Supporting Information, we show that the eigenvector e_l projects the system Eq. (13) onto dynamics of of the center position:

$$\dot{\varphi} = e_l\dot{\delta\mathbf{y}} = e_l K\delta\mathbf{y} = e_l \cdot 0 \cdot \delta\mathbf{y}. \tag{16}$$

Under the linearized ring-attractor dynamics K, the center position is thus not subject to any dynamics, making it susceptible to any displacements by noise.

Calculation of the left eigenvector e_l

If the matrix K is symmetric, the left and right eigenvectors e_l and e_r for the same eigenvalue 0 are the transpose of each other. Unfortunately, here this is not the case (see Eq. (13)), and we need to compute the unknown vector e_l, which will depend on the coefficients of the known vector e_r. In particular, we look for a parametrized vector $\mathbf{y}'(\mathbf{y}) = \left(\mathbf{t}^T(\mathbf{y}), \mathbf{v}^T(\mathbf{y}), \mathbf{z}^T(\mathbf{y})\right)^T$ that for $\mathbf{y} = e_r$ fulfills the transposed eigenvalue equation of the left eigenvector:

$$K^T\mathbf{y}'(e_r) = 0. \tag{17}$$

In Section 2 of S1 Supporting Information, we derive variables $\mathbf{y}'$ that fulfill the transposed dynamics $\dot{\mathbf{y}}' = K^T\mathbf{y}'$ and for which it holds that $\dot{\mathbf{y}}'(e_r) = 0$, thus fulfilling the condition Eq. (17). In this case we know that (due to uniqueness of the 1-dimensional eigenspace associated to the 0 eigenvalue) the vector $\mathbf{y}'^T$ is proportional to e_l:

$$e_l = \frac{1}{S}\mathbf{y}'(e_r)^T = \left(\frac{d\mathbf{J}_0}{d\varphi}^T, \left(\alpha_1\frac{d\mathbf{u}_0}{d\varphi} + \alpha_2\frac{d\mathbf{x}_0}{d\varphi}\right)^T, \left(\beta_1\frac{d\mathbf{u}_0}{d\varphi} + \beta_2\frac{d\mathbf{x}_0}{d\varphi}\right)^T\right), \tag{18}$$

where S is a proportionality constant and $\frac{dJ_{0,i}}{d\varphi} = \sum_j w_{ij}\frac{ds_{0,j}}{d\varphi}$ is the change of the steady-state input arriving at neuron i under shifts of the center position φ.

Finally, the proportionality constant S can be calculated by using Eq. (18) in Eq. (15) (see Section 3 of S1 Supporting Information for details):

$$\begin{aligned}
S =& \mathbf{y}'(e_r)^T \cdot e_r \\
=& U\sum_i \frac{\left(\frac{dJ_{0,i}}{d\varphi}\right)^2 \phi_i'}{\left[U\phi_{0,i}\left(\tau_u\left(\tau_x\phi_{0,i}+1\right)+\tau_x\right)+1\right]^3} \\
& \Big[\tau_s\left[\tau_u\phi_{0,i}\left(U\tau_u\phi_{0,i}+2\right)+1\right]\left[U\phi_{0,i}\left(\tau_u\left(\tau_x\phi_{0,i}+1\right)+\tau_x\right)+1\right] \\
& -\phi_{0,i}\left[\left(U-1\right)\tau_u^2 + U\tau_x^2\left(\tau_u\phi_{0,i}+1\right)\left(\tau_u\phi_{0,i}\left(U\tau_u\phi_{0,i}+2\right)+1\right)\right] \\
& -\frac{(U-1)U\tau_u^2\tau_x\phi_{0,i}\left(\tau_u\phi_{0,i}+1\right)}{\left(U\tau_u\phi_{0,i}+1\right)}\Big],
\end{aligned} \tag{19}$$

where $\phi_{0,i}' = \left.\frac{d\phi_i}{dJ_i}\right|_{J_{0,i}}$ is the linear change of the firing rate of neuron i at its steady-state input $J_{0,i}$.

Diffusion

To be able to describe diffusion on the continuous attractor, we need to extend the model by a treatment of the noise induced into the system through the variable process of neuronal spike emission. Starting from Eq. (3), we assume that neurons i fire according to independent Poisson processes $\xi_i(t) = \sum_k \delta(t - t_{i,k})$, where $t_{i,k}$ is a Poisson point process with time-dependent rate ϕ_i. The variability of the point process $\xi_i(t)$ introduces noise in the synaptic variables. By neglecting the shot-noise (jump-like) nature of this process, we can capture the neurally induced variability simply as white noise with variance proportional to the incoming firing rates [47], $\xi_i(t) = \phi_i + \sqrt{\phi_i}\cdot\eta_i(t)$, where η_i are white Gaussian noise processes with mean $\langle\eta_i\rangle = 0$, and correlation function $\langle\eta_i(t)\eta_j(t')\rangle = \delta(t-t')\delta_{ij}$. This model of $\xi_i(t)$ preserves the mean and the auto-correlation function of the original Poisson processes. Here, we introduce diffusive noise for each synaptic variable separately, but later average their *linear* contributions over the large population, when projecting onto movement along the continuous manifold (see below, and also [39, Supplementary Material] for a discussion).

Substituting the noisy processes $\xi_i(t)$ for $\phi_i(t)$ in Eq. (3) results in the following system of $3\cdot N$ coupled Ito-SDEs:

$$\begin{aligned}
\dot{s}_i &= -\frac{s_i}{\tau_s} + u_i x_i\left(\phi_i + \eta_i\sqrt{\phi_i}\right), \\
\dot{u}_i &= -\frac{u_i - U}{\tau_u} + U(1-u_i)\left(\phi_i + \eta_i\sqrt{\phi_i}\right), \\
\dot{x}_i &= -\frac{x_i - 1}{\tau_x} - u_i x_i\left(\phi_i + \eta_i\sqrt{\phi_i}\right).
\end{aligned} \tag{20}$$

Note that the noise inputs η_i to the synaptic variables for neuron i are all identical, since they result from the same presynaptic spike train.

Linearizing this system around the noise-free steady-state Eq. (12) and considering only the unperturbed noise (we neglect multiplicative noise terms by replacing the terms $\sqrt{\phi_i} \to \sqrt{\phi_{0,i}}$), we arrive at the linearized system equivalent of Eq. (20):

$$\dot{\delta\mathbf{y}} = K\delta\mathbf{y} + \begin{pmatrix} \vec{\eta}\mathbf{u}_0\mathbf{x}_0\sqrt{\vec{\phi}_0} \\ \vec{\eta}U(1-\mathbf{u}_0)\sqrt{\vec{\phi}_0} \\ -\vec{\eta}\mathbf{u}_0\mathbf{x}_0\sqrt{\vec{\phi}_0} \end{pmatrix} \equiv K\delta\mathbf{y} + L. \tag{21}$$

Note that the same vector of white noises $\vec{\eta} \equiv (\eta_1, \ldots, \eta_n)^T$ appears three times.

Left-multiplying this system with the eigenvector e_l yields a stochastic differential equation for the center position (cf. Eq. (16)):

$$\dot{\varphi} = e_l\dot{\delta\mathbf{y}} = 0 \cdot K + e_l L = \sum_k e_{l,k} L_k \tag{22}$$

Through the normalization by S (Eq. (18)), which sums over all neurons, the individual contributions $e_{l,k}$ become small as the number of neurons N increases (this scaling is made explicit in *System size scaling*). Thus, for large networks we average the small contributions of many single noise sources, which validates the diffusion approximation above.

In Section 4 of S1 Supporting Information, we show that we can rewrite Eq. (22) by introducing a single Gaussian white noise process with intensity B (Eq. (5) of the main text), that matches the correlation function of the summed noises:

$$\dot{\varphi} = \sqrt{B}\eta, \tag{23}$$

where η is a white Gaussian noise process with $\langle\eta\rangle = 0$ and $\langle\eta(t)\eta(t')\rangle = \delta(t-t')$. Note, that the value of B is the same under changes of the center position φ: these correspond to index-shifting ($\mod N$) all vectors in Eq. (5), which leaves the sum invariant.

Drift

While the linearization and diffusion coefficient calculated above are invariant with respect to shifts of the bump center, the directed drift introduced by frozen variability will depend on the bump center position φ. Let the center position of the bump be $\varphi_k = k\frac{2\pi}{N} - \pi$ (for $0 \leq k < N$) with the corresponding firing rate profile $\vec{\phi}_0(\varphi_k)$, which is an index-shifted version of the $\phi_{0,i}(\varphi_k) = \phi_{0,i-k}(\varphi_0)$.

To calculate the directed drift for a given position φ_k, we first rotate the system such that the steady-state firing rate profile is again centered at $\varphi_0 = -\pi$. Then, we assume that the effect of all sources of frozen variability can be expressed by a vector of small firing rate perturbations $\Delta\vec{\phi}(\varphi_k)$ to the steady-state firing rate profile:

$$\vec{\phi}_0 \to \vec{\phi}_0 + \Delta\vec{\phi}(\varphi_k). \tag{24}$$

These firing rate perturbations stem from any deviation of the neural system from the "baseline" case and change with each center position φ_k of the bump. The resulting drift field will thus depend on the center position around which we linearize the synaptic dynamics in the following. In *Frozen noise* we will calculate the perturbations induced by random network connectivity, as well as heterogeneous leak reversal-potentials in excitatory neurons of the spiking network.

The firing rate perturbations Eq. (24) add an additional term in the linearized equations Eq. (21):

$$\delta\dot{\mathbf{y}} = K\delta\mathbf{y} + \begin{pmatrix} \mathbf{x}_0\mathbf{u}_0\Delta\vec{\phi}(\varphi_k) \\ U(1-\mathbf{u}_0)\Delta\vec{\phi}(\varphi_k) \\ -\mathbf{x}_0\mathbf{u}_0\Delta\vec{\phi}(\varphi_k) \end{pmatrix} + L. \tag{25}$$

As before, we left-multiply by the left eigenvector e_l, thereby projecting the dynamics onto changes of the center position. This eliminates the linear response kernel K and yields a drift-term in the formerly purely diffusive SDE Eq. (23) (see Section 5 of S1 Supporting Information for details):

$$\dot{\varphi_k} = \sum_i \frac{dJ_{0,i}}{d\varphi}\frac{1}{S}\frac{U\left(1+2\tau_u\phi_{0,i}+U\tau_u^2\phi_{0,i}^2\right)}{\left(U\phi_{0,i}\left(\tau_u\tau_x\phi_{0,i}+\tau_u+\tau_x\right)+1\right)^2}\Delta\phi_i(\varphi_k) + \sqrt{B}\eta. \tag{26}$$

We again assume that the number of neurons N is large enough to treat the center position as a continuous variable $\varphi \in [-\pi, \pi)$, allowing us to treat the set of point-wise linearizations of the dynamics at φ_k as a drift-field $A(\varphi)$ in Eq. (7) of the main text.

For this, it is important to note that, even for random sources of variability, the first term in the point-wise linearization Eq. (26) at each φ_k will vary nearly continuously with changes in the center position. Intuitively, the sum weighs the vector $\Delta\vec{\phi}(\varphi_k)$ of random numbers with a smooth function of the smoothly varying firing-rate profile $\vec{\phi}_0$. Shifts in the center position φ_k yield (to first order) index-shifts in the vector of random numbers (see *Frozen noise*), equivalent to index-shifts of the firing-rate profile. Thus, small changes in center positions will lead to small changes in the summands of Eq. (26). While our results validate the approach, a more rigorous proof of these arguments will be left for future work.

Spiking network model

Spiking simulations are based on a variation of a popular ring-attractor model of visuospatial working memory of [11] (and used with variations in [27, 29, 32, 36, 46]). The recurrent excitatory connections of the original network model have been simplified, to allow for faster simulation as well as analytical derivations of the recurrent synaptic activation. The implementation details are given below, however the major changes are: 1) all recurrent excitatory conductances are voltage independent; 2) a model of synaptic short-term plasticity via facilitation and depression [48, 91, 92] is used to dynamically regulate the weights of the incoming spike-trains 3) recurrent excitatory conductances are computed as linear filters of the weighted incoming spike trains instead of the second-order kinetics for NMDA saturation used in [11].

Neuron model

Neurons are modeled by leaky integrate-and-fire dynamics with conductance based synaptic transmission [11, 49]. The network consists of recurrently connected populations of N_E excitatory and N_I inhibitory neurons, both additionally receiving external spiking input with spike times generated by N_{ext} independent, homogeneous Poisson processes, with rates ν_{ext}. We assume that external excitatory inputs are mediated by fast AMPA receptors, while, for simplicity, recurrent excitatory currents are mediated only by slower NMDA channels.

The dynamics of neurons in both excitatory and inhibitory populations are governed by the following system of differential equations indexed by $i \in \{0, ..., N_{E/I} - 1\}$:

$$\begin{aligned} C_m \dot{V}_i(t) &= -I_i^{\mathrm{L}}(t) - I_i^{\mathrm{Ext}}(t) - I_i^{\mathrm{I}}(t) - I_i^{\mathrm{E}}(t), \\ I_i^P &= g_P \, s_i^P(t) \, (V_i(t) - V_P) \, , \end{aligned} \tag{27}$$

where $P \in \{\mathrm{L,Ext,I,E}\}$, V denotes voltages (membrane potential) and I denotes currents. Here, C_{m} is the membrane capacitance and $V_{\mathrm{L}}, V_{\mathrm{E}}, V_{\mathrm{I}}$ are the reversal potentials for leak, excitatory currents, and inhibitory currents, respectively. The parameters g_P for $P \in \{\mathrm{L,Ext,I,E}\}$ are fixed scales for leak (L), external input (Ext) and recurrent excitatory (E) and inhibitory (I) synaptic conductances, which are dynamically gated by the unit-less gating variables $s_i^P(t)$. These gating variables are described in detail below, however we set the leak conductance gating variable to $s_i^{\mathrm{L}} = 1$. For excitatory neurons, we refer to the excitatory and inhibitory conductance scales by $g_{\mathrm{EE}} \equiv g_{\mathrm{E}}$ and $g_{\mathrm{EI}} \equiv g_{\mathrm{I}}$, respectively. Similarly, for inhibitory neurons, we refer to the excitatory and inhibitory conductance scales by $g_{\mathrm{IE}} \equiv g_{\mathrm{E}}$ and $g_{\mathrm{II}} \equiv g_{\mathrm{I}}$, respectively.

The model neuron dynamics (Eq. 27) are integrated until their voltage reaches a threshold V_{thr}. At any such time, the respective neuron emits a spike and its membrane potential is reset to the value V_{res}. After each spike, voltages are clamped to V_{res} for a refractory period of τ_{ref}. See the Tables in S1 Table and S2 Table for parameter values used in simulations.

Synaptic gating variables and short-term plasticity

The unit-less synaptic gating variables $s_i^P(t)$ for $P \in \{\mathrm{Ext,I}\}$ (external and inhibitory currents) are exponential traces of the spike trains of all presynaptic neurons j with firing times t_j:

$$\dot{s}_i^P(t) = -\frac{s_i^P(t)}{\tau_P} + \sum_{j \in \mathrm{pre}(P)} w_{ij}^P \sum_{t_j} \delta\left(t - t_j\right), \tag{28}$$

where pre(P) indicates all neurons presynaptic to the neuron i for the the connection type P. The factors w_{ij}^P are unit-less synaptic efficacies for the connection from neuron j to neuron i. For the excitatory gating variables of inhibitory neurons s_i^{IE} (IE denotes connections from E to I neurons) we also use the linear model of Eq. (28) with time constant $\tau_{IE} = \tau_E$.

For excitatory to excitatory conductances, we use a well established model of synaptic short-term plasticity (STP) [48, 91, 92] which provides dynamic scaling of synaptic efficacies depending on presynaptic firing. This yields two additional dynamical variables, the facilitating synaptic efficacy $u_j(t)$, as well as the fraction of available synaptic resources $x_j(t)$ of the outgoing connections of a presynaptic neuron j, which are implemented according to the following differential equation:

$$\begin{aligned} \dot{u}_j &= -\frac{1}{\tau_u}\left(u_j - U\right) + U\left(1 - u_j^-\right) \sum_{t_j} \delta(t - t_j), \\ \dot{x}_j &= -\frac{1}{\tau_x}\left(x_j - 1\right) - x_j^- u_j^- \sum_{t_j} \delta(t - t_j). \end{aligned} \tag{29}$$

Here, the indices u_j^- and x_j^- indicate that for the incremental update of the variables upon spike arrival, the values of the respective variables immediately before the spike arrival are used [92]. The variable U appears in the equation for $u(t)$ both as the steady-state value in the absence of spikes and as a scale for the per-spike – intuitively, U is thus the first value of $u(t)$transmitted on a spike after inactivity of the connection [48].

The dynamics of recurrent excitatory-to-excitatory transmission with STP are then given by gating variables that linearly filter the incoming spikes, which are scaled by facilitation and depression:

$$s_i^{\mathrm{EE}}(t) = \sum_{j \in \mathrm{pre(EE)}} w_{ij}^{\mathrm{EE}} s_j \tag{30}$$

$$\dot{s}_j = -\frac{s_j}{\tau_s} + \sum_{t_j} \delta\left(t - t_j\right) u_j^-(t) x_j^-(t). \tag{31}$$

Here, pre(EE) indicates all excitatory neurons that make synaptic connections to the neuron i. See the Table in S2 Table for synaptic parameters used in simulations.

This makes the system Eqs. (29) and (31) a spiking variant of the rate-based dynamics of Eq. (3), and s_i^{EE} a variable related to the input J_i (cf. Eq. (2)). In *Firing rate approximation* we will make this link explicit.

Network connectivity

All connections except for the recurrent excitatory connections are all-to-all and uniform, with unit-less connection strengths set to $w_{ij}^{\mathrm{I}} = w_{ij}^{\mathrm{ext}} = 1$ and for inhibitory neurons additionally $w_{ij}^{\mathrm{E}} = 1$. The recurrent excitatory connections are distance-dependent and symmetric. Each neuron of the excitatory population with index $i \in \{0, ..., N_E - 1\}$ is assigned an angular position $\theta_i = i \cdot \frac{2\pi}{N_E} \in [0, 2\pi)$. Recurrent excitatory connection weights w_{ij}^{EE} from neuron j to neuron i are then given by the Gaussian function $w^{\mathrm{EE}}(\theta)$ as (see the Table in S2 Table for parameters used in simulations):

$$\begin{aligned} w_{ij}^{\mathrm{EE}} &= w^{\mathrm{EE}}(\theta_i - \theta_j) \\ &= w_0 + (w_+ - w_0) \exp\left(- \left[\min\left(|\theta_i - \theta_j|, 2\pi - |\theta_i - \theta_j|\right)\right]^2 \frac{1}{2\sigma_w^2} \right). \end{aligned} \tag{32}$$

Additionally, for each neuron we keep the integral over all recurrent connection weights normalized, resulting in the normalization condition $\frac{1}{2\pi}\int_{-\pi}^{\pi} d\varphi w^{\mathrm{EE}}(\varphi) = 1$. This normalization ensures that varying the maximum weight w_+ will not change the total recurrent excitatory input if all excitatory neurons fire at the same rate. Here, we choose w_+ as a free parameter constraining the baseline connection weight to:

$$w_0 = \frac{w_+ \sigma_w \mathrm{erf}\left(\frac{\pi}{\sqrt{2}\sigma_w}\right) - \sqrt{2\pi}}{\sigma_w \mathrm{erf}\left(\frac{\pi}{\sqrt{2}\sigma_w}\right) - \sqrt{2\pi}}.$$

Firing rate approximation

We first replace the synaptic activation variables $s^P(V, t)$ for $P \in \{\mathrm{I}, \mathrm{ext}\}$ by their expectation values under input with Poisson statistics. We assume that the inhibitory population fires at rates ν_I. For the linear synapses this yields

$$\left\langle s^{\mathrm{ext}} \right\rangle = \tau_{\mathrm{ext}} N_{\mathrm{ext}} \nu_{\mathrm{ext}}, \tag{33}$$

$$\left\langle s^{\mathrm{I}} \right\rangle = \tau_{\mathrm{I}} N_I \nu_{\mathrm{I}}. \tag{34}$$

For the recurrent excitatory-to-excitatory synapses with short-term plasticity, we set the differential equations (29) to zero, and also average them over the Poisson statistics. Akin to the "mean-field" model of [48], we average the steady-state values of facilitation and depression separately over the Poisson statistics. This implicitly assumes that facilitation and depression are statistically independent, with respect to the distributions

of spike times – while this is not strictly true, the approximations work well, as has been previously reported [48]. This allows a fairly straightforward evaluation of the mean steady-state value of the combined facilitation and depression variables $\langle u_j x_j \rangle$, under the assumption that the neuron j fires at a mean rate ν_j with Poisson statistics, and yields rate approximations of the steady-state values similar to Eq. (12):

$$\langle u_j x_j \rangle = \langle u_j \rangle \langle x_j \rangle = \frac{U\left(\nu_j \tau_u + 1\right)}{U \nu_j (\tau_u + \tau_x + \nu_j \tau_u \tau_x) + 1}. \tag{35}$$

We now assume that the excitatory population of N_E neurons fires at the steady-state rates ϕ_j $(0 \leq j < N)$. To calculate the synaptic activation of excitatory-to-excitatory connections $\langle s_i^{\mathrm{EE}} \rangle$, we set Eq. (30) to zero, and average over Poisson statistics (again neglecting correlations), which yields$\langle s_j \rangle = \tau_E \langle u_j x_j \rangle \phi_j$ and $\langle s_i^{\mathrm{EE}} \rangle = \sum_j w_{ij}^{\mathrm{EE}} \tau_E \langle u_j x_j \rangle \phi_j$. Let the the normalized steady-state input J_i be:

$$J_i \equiv \frac{1}{N_{\mathrm{E}}} \langle s_i^{\mathrm{EE}} \rangle = \frac{1}{N_{\mathrm{E}}} \sum_j w_{ij}^{\mathrm{EE}} \langle s_j \rangle. \tag{36}$$

The steady-state input Eq. (36) links the general framework of Eq. (2) to the spiking network. The additional factor $1/N_E$ is introduced to make the scaling of the excitatory-to-excitatory conductance with the size of the excitatory population N_{E} explicit, which will be used in *System size scaling*. To see this, we assume that the excitatory conductance scale of excitatory neurons g_{EE} is scaled such that the total conductance is invariant under changes of N_{E} [93]: $g_{\mathrm{EE}} = \tilde{g}_{\mathrm{EE}}/N_E$, for some fixed value $\tilde{g}_{\mathrm{EE}}$. This yields the total excitatory-to-excitatory conductance $g_{\mathrm{EE}} s_i^{\mathrm{EE}} = \tilde{g}_{\mathrm{EE}} J_i$ with J_i as introduced above, where the scaling with N_{E} is now shifted to the input variable J_i.

For the synaptic activation of excitatory to inhibitory connections, we get the mean activations:

$$\left\langle s^{\mathrm{IE}} \right\rangle = \tau_{\mathrm{E}} \sum_j \phi_j. \tag{37}$$

We then follow [49] to reduce the differential equations of Eq. (27) to a dimensionless form. The main difference consists in the absence of the voltage dependent NMDA conductance, which is achieved by setting the two associated parameters $\beta \to 0, \gamma \to 0$ in [49], to arrive at:

$$\begin{aligned}
\tau_i \dot{V}_i &= -\left(V_i - V_L\right) + \mu_i + \sigma_i \sqrt{\tau_i} \eta_i(t) \\
S_i &= 1 + T_I \nu_I + T_{\mathrm{ext}} \nu_{\mathrm{ext}} + T_E J_i && (38) \\
\mu_i S_i &= \left(V_I - V_L\right) T_I \nu_I + \left(V_E - V_L\right) T_{\mathrm{ext}} \nu_{\mathrm{ext}} + \left(V_E - V_L\right) T_E J_i && (39) \\
\sigma_i &= \frac{g_{\mathrm{ext}}}{C_m} \left(\langle V \rangle - V_E\right) \tau_{\mathrm{ext}} \sqrt{\tau_i N_{\mathrm{ext}} \nu_{\mathrm{ext}}}. && (40) \\
\tau_i &= \frac{C_m}{g_L S_i} \\
\langle \eta_i(t) \rangle &= 0 \\
\langle \eta_i(t) \eta_i(t') \rangle &= \frac{1}{\tau_{\mathrm{ext}}} \exp\left(-\frac{|t - t'|}{\tau_{\mathrm{ext}}}\right) && (41)
\end{aligned}$$

where $T_{\mathrm{ext}} = N_{\mathrm{ext}} \tau_{\mathrm{ext}} \frac{g_{\mathrm{ext}}}{g_L}$, $T_I = N_I \tau_I \frac{g_I}{g_L}$ are effective timescales of external and inhibitory inputs, and $T_E = N_E \frac{g_E}{g_L}$ is a dimensionless scale for the excitatory conductance. Here, μ_i is the bias of the membrane potential due to synaptic inputs, and σ_i measures the scale of fluctuations in the membrane potential due to random spike arrival approximated by the Gaussian process η_i.

The mean firing rates F and mean voltages $\langle V_i \rangle$ of populations of neurons governed by this type of differential equation can then be approximated by:

$$
\begin{aligned}
F\left[\mu_i, \sigma_i, \tau_i\right] &= \left(\tau_{\mathrm{ref}} + \sqrt{\pi}\tau_i \int_{\beta(\mu_i,\sigma_i)}^{\alpha(\mu_i,\sigma_i)} du \exp(u^2)\left[1 + \mathrm{erf}\left(u\right)\right]\right)^{-1} && (42)\\
\alpha(\mu_i, \sigma_i, \tau_i) &= \frac{V_{\mathrm{reset}} - V_L - \mu_i}{\sigma_i}\left(1 + \frac{\tau_{\mathrm{ext}}}{2\tau_i}\right) + 1.03\sqrt{\frac{\tau_{\mathrm{ext}}}{\tau_i}} - \frac{\tau_{\mathrm{ext}}}{\tau_i}, && (43)\\
\beta(\mu_i, \sigma_i) &= \frac{V_{\mathrm{reset}} - V_L - \mu_i}{\sigma_i}, && (44)\\
\langle V_i \rangle &= \mu_i + V_L - \left(V_{\mathrm{thr}} - V_{\mathrm{reset}}\right)\phi_i\tau_i. && (45)
\end{aligned}
$$

Derivatives of the rate prediction

Here we calculate derivatives of the input-output relation (Eq. (42)) that will be used below in *Frozen noise*.

The expressions for drift and diffusion (see *Analysis of drift & diffusion with STP*) contain the derivative $\phi_i' = \frac{dF}{dJ}\big|_{J_i}$ of the input-output relation F (Eq. (42)) with respect to the recurrent excitatory input J_i. Note, that F depends on J_i through all three arguments μ_i, σ_i and τ_i. First, we define $X(u) \equiv \exp(u^2)\left[1 + \mathrm{erf}\left(u\right)\right]$, and the shorthand $F_i = F\left[\mu_i, \sigma_i, \tau_i\right]$. The derivative can then be readily evaluated as (to shorten the notation in the following, we skip noting the evaluation points for derivatives in the following):

$$
\begin{aligned}
\frac{dF}{dJ_i} &= -F_i\frac{T_E}{S_i}\left(F_i\tau_{\mathrm{ref}} - 1\right) + \sqrt{\pi}F_i^2\tau_i\left[X(\beta)\frac{d\beta}{dJ_i} - X(\alpha)\frac{d\alpha}{dJ_i}\right], && (46)\\
\frac{d\alpha/\beta}{dJ_i} &= \left(\frac{\partial\alpha/\beta}{\partial\mu_i} + \frac{\partial\alpha/\beta}{\partial\sigma_i}\frac{\partial\sigma_i}{\partial\langle V_i\rangle}\right)\frac{T_E}{S_i}\left(-\mu_i + (V_E - V_L)\right)\\
&\quad - \left(\frac{\partial\alpha/\beta}{\partial\sigma_i}\left[\frac{\partial\sigma_i}{\partial\tau_i} - \frac{\partial\sigma_i}{\partial\langle V_i\rangle}\left(V_{\mathrm{thr}} - V_{\mathrm{reset}}\right)\phi_i\right] + \frac{\partial\alpha/\beta}{\partial\tau_i}\right)\frac{T_E}{S_i}\tau_i.
\end{aligned}
$$

where α/β stands as a placeholder for either function, and the expressions for α and β are given in Eqs. (43)-(44).

A second expression involving the derivative of Eq. (42) is $\frac{d\phi_{0,i}}{d\Delta_i^{\mathrm{L}}}$ which appears in the theory when estimating firing rate perturbations caused by frozen heterogeneities in the leak potentials of excitatory neurons (see Eq. (54)). The resulting derivatives are almost similar, which can be seen by the fact that replacing $V_{\mathrm{L}} \to V_{\mathrm{L}} + \Delta_i^{\mathrm{L}}$ in Eq. (27) only leads to an additional term Δ_i^{L} in Eq. (39). Thus, for neuron i the derivative can be evaluated to

$$
\begin{aligned}
\frac{dF}{d\Delta_i^L} &= \sqrt{\pi}F_i^2\tau_i\left[X(\beta)\frac{\partial\beta}{\partial\Delta_i^{\mathrm{L}}} - X(\alpha)\frac{\partial\alpha}{\partial\Delta_i^{\mathrm{L}}}\right], && (47)\\
\frac{d\alpha/\beta}{d\Delta_i^{\mathrm{L}}} &= \left(\frac{\partial\alpha/\beta}{\partial\mu_i} + \frac{\partial\alpha/\beta}{\partial\sigma_i}\frac{\partial\sigma_i}{\partial\langle V_i\rangle}\right)\frac{1}{S_i}.
\end{aligned}
$$

In practice, given a vector $\phi_{i,0}$ of firing rates in the attractor state, as well as the mean firing rate of inhibitory neurons ν_I, we evaluate the right hand side of Eq. (46) and Eq. (47) by replacing $F_i \to \phi_{i,0}$. This allows efficiently calculating the derivatives without having to perform any numerical integration. The two terms will be exactly equal if $\phi_{0,i}$ is a self-consistent solution of Eq. (42) for firing rates of the excitatory neurons across the network. We used numerical estimates of $\phi_{i,0}$ and ν_I that were measured from simulations and were very close to firing-rate predictions for all networks we investigated.

Optimization of network parameters

We used an optimization procedure [94] to retune network parameters to produce approximately similar bump shapes as the parameters of short-term plasticity are varied. Briefly, we replace the network activity ϕ_j in the total input J_i of Eq. (36) by a parametrization

$$g(\theta_j) = g_0 + g_1 \exp\left(-\left[\frac{|\theta_j|}{g_\sigma}\right]^{g_r}\right). \tag{48}$$

Approximating sums $\frac{1}{N_E}\sum_{j=0}^{N_E-1}$ with integrals $\frac{1}{2\pi}\int_{-\pi}^{\pi} d\varphi$ we arrive at

$$J_i(g) \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} d\varphi w^{\mathrm{EE}}\left(\theta_i - \varphi\right)\langle s_j\rangle(g\left(\varphi\right)),$$

where $J_i(g)$ indicates that the total input depends on the parameters g_0, g_1, g_σ, g_r of the parametrization g.

We then substitute this relation in Eq. (42) to arrive at a self-consistency relation between the parametrized network activity $g(\theta_i)$ at the position of neuron i and the firing-rate F predicted by the theory:

$$g(\theta_i) = F\left[\mu_i(g), \sigma_i(g), \tau_i(g), \langle V_i\rangle\left(g\right)\right]. \tag{49}$$

As for the input $J_i(g)$, we indicate the dependence of quantities on the parameters of the parametrization g by appending (g). The explicit dependence of the voltage $\langle V_i\rangle\left(g\right)$ on g is obtained by additionally substituting $\phi_i \to g(\theta_i)$ in Eq. (45).

We then optimized networks to fulfill Eq. (49). First, we imposed the following targets for the parameters of g: $g_0 = 0.1$Hz, $g_1 = 40$Hz, $\nu_{E,\mathrm{basal}} = 0.5$Hz, $\nu_{I,\mathrm{basal}} = 3$Hz. For all networks we chose $w_+ = 4.0$, $g_r = 2.5$. The following parameters were then optimized: ν_I, g_σ, g_{EE} (excitatory conductance g_E on excitatory neurons); g_{IE} (excitatory conductance g_E on inhibitory neurons); g_{EI} (inhibitory conductance g_I on excitatory neurons); g_{II} (inhibitory conductance g_I on inhibitory neurons). The basal firing rates (firing rates in the uniform state of the network, prior to being cued) yielded two equations from Eq. (49) by setting $w_+ = 1$. This left 4 free parameters, which were constrained by evaluating Eq. (49) at 4 points as described in [94]. The basal firing rates were chosen to be fairly low to make the uniform state more stable (as in [44]). This procedure does not yield a fixed value for g_σ, since g_σ is optimized for and is not set as a target value. We thus iterated the following until a solution was found with $g_\sigma \approx 0.5$: a) change the width of the recurrent weights w_σ; b) optimize network parameters as described here; c) optimize the expected bump shape for the new network parameters to predict g_σ. The resulting parameter values are given in Table in S2 Table.

Frozen noise

Random and heterogeneous connectivity

Introducing random connectivity, we replace the recurrent weights in Eq. (36) by:

$$w_{ij}^{\mathrm{EE}} \quad \to \quad \left[w_{ij}^{\mathrm{EE}} + \Delta_{ij}^{w}\right]\frac{p_{ij}}{p}. \tag{50}$$

Here, $p_{ij} \in \{0, 1\}$ are Bernoulli variables, with $P(p_{ij} = 1) = p$, where the connectivity parameter $p \in (0, 1]$ controls the overall sparsity of recurrent excitatory connections. For $p = 1$ the entire network is all-to-all connected. Additionally, we provide derivations for additive synaptic heterogeneities $\Delta_{ij}^{w} = \eta_{ij}\sigma_w$ (as in [38]), where $\{\eta_{ij}|1 \le i, j \le N_E\}$ are independent, normally distributed random variables with zero mean and unit variance.

We did not investigate this type of heterogeneity in the main text, since increasing σ_w lead to a loss of the attractor state before creating large enough directed drifts to be comparable to the other sources of frozen noise considered here – most of the small effects were "hidden" behind diffusive displacement [81]. Nevertheless, we included this case in the analysis here for completeness.

Let the center position of the bump be $\varphi_k = k\frac{2\pi}{N} - \pi$ (for $0 \leq k < N$). Subject to the perturbed weights, the recurrent steady-state excitatory input $J_i(\varphi_k)$ Eq. (36) to any excitatory neuron can be written as the unperturbed input $J_{0,i}(\varphi_k)$ plus an additional input $J_i^{\text{struct}}(\varphi_k)$ arising from the perturbed connectivity. Note that the synaptic steady-state activations $s_{0,j}(\varphi_k)$ change with varying bump centers – in the following, we denote $s_{0,j}^k \equiv s_{0,j}(\varphi_k)$:

$$\begin{aligned}
J_i(\varphi_k) &= \frac{1}{N_E}\sum_j \left[w_{ij}^{\text{EE}} + \Delta_{ij}^w\right]\frac{p_{ij}}{p}s_{0,j}^k \\
&= \frac{1}{N_E}\frac{1}{p}\left[\sum_j \left[w_{ij}^{\text{EE}} + \Delta_{ij}^w\right]s_{0,j}^k - \sum_j \left[w_{ij}^{\text{EE}} + \Delta_{ij}^w\right](1-p_{ij})\,s_{0,j}^k\right] \\
&= \underbrace{\frac{1}{N_E}\sum_j w_{ij}^{\text{EE}}s_{0,j}^k}_{J_{0,i}(\varphi_k)} + \underbrace{\frac{1}{N_E}\frac{1}{p}\left[\sum_j \left[w_{ij}^{\text{EE}} + \Delta_{ij}^w\right]p_{ij}s_{0,j}^k - p\sum_j w_{ij}^{\text{EE}}s_{0,j}^k\right]}_{J_i^{\text{struct}}(\varphi_k)}.
\end{aligned}$$

Note that $J_{0,i}(\varphi_k)$ is an index-shifted version of the steady-state input: $J_{0,i}(\varphi_k) = J_{0,i-k}$. However, such a relation does not hold for $J_i^{\text{struct}}(\varphi_k)$, since the random numbers p_{ij} will change the resulting value for varying center positions.

We calculate the firing rate perturbations $\delta\phi_i(\varphi_k)$ resulting from the additional input by a linear expansion around the steady-state firing rates $\phi_{0,i}(\varphi_k) \to \phi_{0,i}(\varphi_k) + \delta\phi_i(\varphi_k)$. These evaluate to:

$$\begin{aligned}
\delta\phi_i(\varphi_k) &= \left.\frac{dF}{dJ}\right|_{J_{0,i}}(\varphi_k)\cdot J_i^{\text{struct}}(\varphi_k) \\
&= \phi'_{i,0}(\varphi_k)\cdot J_i^{\text{struct}}(\varphi_k).
\end{aligned} \tag{51}$$

See *Derivatives of the rate prediction* for the derivation of the function $\frac{dF}{dJ}(J_{0,i})$ for the spiking network used in the main text.

In the sum of Eq. (7), we keep the firing rate profile $\vec{\phi}_0$ centered at φ_0 while calculating the drift for varying center positions. To accommodate the shifted indices resulting from moving center positions, we re-index the summands to yields the perturbations $\phi_{0,i} \to \phi_{0,i} + \Delta\phi_i(\varphi_k)$ used there:

$$\Delta\phi_i(\varphi_k) = \phi'_{i,0}\cdot J_{i+k}^{\text{struct}}(\varphi_k). \tag{52}$$

Heterogeneous leak reversal potentials

We further investigated random distributions of the leak reversal potential V_{L}. These are implemented by the substitution

$$V_{\text{L}} \quad\to\quad V_{\text{L}} + \Delta_i^{\text{L}}, \tag{53}$$

where the Δ_i^{L} are independent normally distributed variables with zero mean, i.e. $\left\langle\Delta_i^{\text{L}}\right\rangle = 0mV, \left\langle\Delta_i^{\text{L}}\Delta_j^{\text{L}}\right\rangle = \sigma_{\text{L}}^2\delta_{ij}$. The parameter σ_{L} controls the standard deviation of these random variables, and thus the noise level of the leak heterogeneities.

Let $\varphi_k = k\frac{2\pi}{N} - \pi$ for $0 \leq k < N$ be the center position of the bump. First, note that the heterogeneities Δ_i^L do not depend on the center position φ_k, since they are single neuron properties. As in the last section, we calculate the firing rate perturbations $\delta\phi_i(\varphi_k)$ resulting from the additional input by a linear expansion around the steady-state firing rates $\phi_{0,i}(\varphi_k) \to \phi_{0,i}(\varphi_k) + \delta\phi_i(\varphi_k)$:

$$\begin{aligned}\delta\phi_i(\varphi_k) &= \frac{dF_i}{d\Delta_i^{\mathrm{L}}}(J_i(\varphi_k)) \cdot \Delta_i^{\mathrm{L}} \\ &\equiv \frac{d\phi_{0,i}}{d\Delta_i^{\mathrm{L}}}(\varphi_k) \cdot \Delta_i^{\mathrm{L}}. \end{aligned} \tag{54}$$

Here, $\frac{dF_i}{d\Delta_i^{\mathrm{L}}}(J_i(\varphi_k))$ is the derivative of the input-output relation of neuron i in a bump centered at φ_k, with respect to the leak perturbation. We introduced $\frac{d\phi_{0,i}}{d\Delta^L}(\varphi_k)$ as a shorthand notation for this derivative, since it is evaluated at the steady-state input $J_{i,0}(\varphi_k)$. For the spiking network of the main text, this is derived in *Derivatives of the rate prediction*.

In the sum of Eq. (7), we keep the firing rate profile $\vec{\phi}_0$ centered at φ_0 while calculating the drift for varying center positions. As in the last section, we re-index the sum to yield the perturbations $\phi_{0,i} \to \phi_{0,i} + \Delta\phi_i(\varphi_k)$ used there:

$$\Delta\phi_i(\varphi_k) = \frac{d\phi_{0,i}}{d\Delta_i^{\mathrm{L}}} \cdot \Delta_{i+k}^{\mathrm{L}}. \tag{55}$$

Squared field magnitude

Using the equation of the drift field in Eq. (7), and the firing rate perturbations Eqs. (51)-(54), it is straight forward to see that for any center position φ the expected drift field averaged over the noise parameters is 0, since all single firing rate perturbations vanish in expectation. In the following we calculate the variance of the drift field averaged over noise realizations, which turns out to be additive with respect to the two noise sources.

We begin by calculating the correlations between frozen noises caused by random connectivity and leak heterogeneities. For the Bernoulli distributed variables p_{ij} it holds that $\langle p_{ij}\rangle = p, \langle p_{ij}p_{lk}\rangle = \delta_{il}\delta_{jk}p + (1-\delta_{il}\delta_{jk})p^2$. For the other independent random variables it holds that $\left\langle \Delta_i^L \right\rangle = 0mV, \left\langle \left(\Delta_i^L\right)^2 \right\rangle = \sigma_L^2, \left\langle \Delta_{ij}^w \right\rangle = 0, \left\langle \left(\Delta_{ij}^w\right)^2 \right\rangle = \sigma_w^2$. Again, the weight heterogeneities Δ_{ij}^w are only included for completeness – all analyses of the main text assume that $\sigma_w = 0$.

For the correlations between the perturbations we then know that (for brevity, we omit the dependence on the center position φ):

$$\begin{aligned}\left\langle J_i^{\mathrm{struct}} \Delta_i^L \right\rangle &= 0 \\ \left\langle J_i^{\mathrm{struct}} J_l^{\mathrm{struct}} \right\rangle &= \frac{1}{N_E^2}\left(\sum_{j,k} s_{0,j} w_{ij}^{\mathrm{EE}} s_{0,k} w_{lk}^{\mathrm{EE}} \left\langle \left(\frac{p_{ij}}{p} - 1\right)\left(\frac{p_{lk}}{p} - 1\right)\right\rangle\right) \\ &\quad + \frac{1}{N_E^2}\left(\frac{1}{p^2}\sum_{j,k} s_{0,j} s_{0,k} \left\langle \Delta_{ij}^w \Delta_{lk}^w \right\rangle \left\langle p_{ij} p_{lk}\right\rangle\right) \\ &= \frac{1}{N_E^2}\left(\sum_j s_{0,j}^2 \left(w_{ij}^{\mathrm{EE}}\right)^2 \left(\frac{1}{p} - 1\right) + \frac{1}{p}\sum_j s_{0,j}^2 \sigma_w^2\right)\delta_{il}. \end{aligned}$$

Starting from Eq. (7), we use as a firing rate perturbation the sum of firing rate perturbations from both Eq. (51) and Eq. (54). With the pre-factor

$C_i = \frac{dJ_{0,i}}{d\varphi} \frac{1+\tau_u \phi_{0,i}(U\tau_u \phi_{0,i}+2)}{(U\phi_{0,i}(\tau_u \tau_x \phi_{0,i}+\tau_u+\tau_x)+1)^2}$, the expected squared field averaged over ensemble of frozen noises is then:

$$\begin{aligned}
\left\langle A(\varphi)^2 \right\rangle_{\text{frozen}} &= \frac{1}{S^2} \sum_{i,j} C_i(\varphi) C_j(\varphi) \left\langle \left(\phi'_{0,i}(\varphi) J_i^{\text{struct}}(\varphi) + \frac{d\phi_{0,i}}{d\Delta_L}(\varphi) \Delta_i^L \right) \cdot \right. \\
&\qquad \left. \left(\phi'_{0,j}(\varphi) J_j^{\text{struct}}(\varphi) + \frac{d\phi_j}{d\Delta_L}(\varphi) \Delta_j^L \right) \right\rangle_{\text{frozen}} \\
&= \frac{1}{S^2} \sum_i C_i^2(\varphi) \cdot \\
&\quad \left[\frac{\left(\phi'_{0,i}(\varphi)\right)^2}{N_E^2} \left(\left(\frac{1}{p} - 1 \right) \sum_j \left(s_{0,j}(\varphi)\right)^2 \left(w_{ij}^{\text{EE}}\right)^2 + \frac{1}{p} \sum_j s_{0,j}^2(\varphi) \sigma_w^2 \right) \right. \\
&\quad \left. + \left(\frac{d\phi_{0,i}}{d\Delta_L}(\varphi) \right)^2 \sigma_L^2 \right]. \qquad (56)
\end{aligned}$$

One can see directly that the two last terms are invariant under shifts of the bump center φ, since these introduce symmetric shifts of the indexes i. Similarly, it is easy to see that the first term is also invariant. Let φ' be shifted to the right by one index from φ . It then holds that:

$$\begin{aligned}
&\sum_i C_i^2(\varphi') \left(\frac{\left(\phi'_{0,i}(\varphi')\right)^2}{N_E^2} \left(\frac{1}{p} - 1 \right) \sum_j \left(s_{0,j}(\varphi')\right)^2 \left(w_{ij}^{\text{EE}}\right)^2 \right) \\
=& \sum_i C_{i-1}^2 \left(\frac{\left(\phi'_{0,i-1}(\varphi)\right)^2}{N_E^2} \left(\frac{1}{p} - 1 \right) \sum_j \left(s_{0,j-1}(\varphi)\right)^2 \left(w_{ij}^{\text{EE}}\right)^2 \right) \\
=& \sum_i C_{i-1}^2 \left(\frac{\left(\phi'_{0,i-1}(\varphi)\right)^2}{N_E^2} \left(\frac{1}{p} - 1 \right) \sum_j \left(s_{0,j}(\varphi)\right)^2 \left(w_{i-1,j}^{\text{EE}}\right)^2 \right).
\end{aligned}$$

The final equation holds since, in ring-attractor networks, w_{ij}^{EE} consists of index-shifted rows of the same vector (see e.g. *Network connectivity* for the spiking network weights).

In summary, $\left\langle A(\varphi)^2 \right\rangle_{\text{frozen}}$ will evaluate to the same quantity $\left\langle A^2 \right\rangle_{\text{frozen}}$ for all center positions φ. In the main text, we use this fact to estimate $\left\langle A^2 \right\rangle_{\text{frozen}}$ from simulations, by additionally averaging over the all center positions and interchanging the ensemble and positional averages:

$$\left\langle A^2 \right\rangle_{\text{frozen}} = \frac{1}{N_E} \sum_k \left\langle A(\varphi_k)^2 \right\rangle_{\text{frozen}} = \left\langle \frac{1}{N_E} \sum_k A(\varphi_k)^2 \right\rangle_{\text{frozen}}.$$

Thus, we can compare the value of $\left\langle A^2 \right\rangle_{\text{frozen}}$ to the mean squared drift field over all center positions, averaged over instantiations of noises.

System size scaling

Generally, sums over the discretized intervals $[-\pi, \pi)$ as they appear in Eqs. (5) and (7) will scale with the number N chosen for the discretization of the positions on the continuous ring $\varphi(i) = \frac{i}{N} 2\pi - \pi$. Consider two discretizations of the ring, partitioned into N_1 and N_2 uniformly spaced bins of width $\frac{2\pi}{N_1}$ and $\frac{2\pi}{N_2}$. We can then approximate

integrals over any continuous (Riemann integrable) function f on the ring by the two Riemann sums:

$$\frac{2\pi}{N_1}\sum_{i=0}^{N_1-1} f(\varphi_{1,i}) \approx \int_{-\pi}^{\pi} f(\varphi)d\varphi \approx \frac{2\pi}{N_2}\sum_{i=0}^{N_2-1} f(\varphi_{2,i}), \tag{57}$$

where, $i\frac{2\pi}{N_1} \leq \varphi_{1,i} < (i+1)\frac{2\pi}{N_1}$ (for N_2 and $\varphi_{2,i}$ analogously) are points in the bins [95].

Numerical quantities for the results of the main text have been calculated for $N_E = 800$. In the following we denote all of these quantities with an asterisk ($*$). To generalize these results to arbitrary system size N, we replace sums over N bins by scaled sums over N_E bins using the relation Eq. (57):

$$\sum_{i=0}^{N-1} \to \frac{N}{N_E}\sum_{i=0}^{N_E-1}.$$

First, we find that the normalization constant scales as $S = \frac{N}{N_E}S^*$, and thus (dots indicate the summands, which are omitted for clarity) for the diffusion strength B (cf. Eq. (5)):

$$\begin{aligned} B &= \frac{1}{S^2}\sum_{i=0}^{N-1}\cdots = \frac{N}{N_E}\frac{1}{S^2}\sum_{i=0}^{N_E-1}\cdots \\ &= \frac{N_E}{N}B^*. \end{aligned} \tag{58}$$

For the drift magnitude we turn to the expected squared drift magnitude calculated earlier (cf. Eq. (56)), for which we find that (setting $\sigma_w \to 0$ for simplicity, as throughout the main text):

$$\begin{aligned} \left\langle A^2\right\rangle_{\text{frozen}} &= \left(\frac{N_E}{N}\right)^2 \frac{1}{(S^*)^2}\frac{N}{N_E}\sum_{i=0}^{N_E-1} C_i^2 \left(\frac{(\phi_i')^2}{N^2}\left(\frac{1}{p}-1\right)\frac{N}{N_E}\sum_{j=0}^{N_E-1} s_j^2 w_{ij}^2 + \left(\frac{d\phi_i}{dE_L}\right)^2\sigma_L^2\right) \\ &= \frac{1}{(S^*)^2}\sum_{i=0}^{N_E-1} C_i^2 \left(\frac{1}{N^2}(\phi_i')^2\left(\frac{1}{p}-1\right)\sum_{j=0}^{N_E-1} s_j^2 w_{ij}^2 + \frac{N_E}{N}\left(\frac{d\phi_i}{dE_L}\right)^2\sigma_L^2\right). \end{aligned} \tag{59}$$

Note, that we could not resolve this scaling in dependence of $\left\langle A^2\right\rangle^*_{\text{frozen}}$, since the two sources of frozen noise (connectivity and leak heterogeneity) show different scaling with N.

Numerical methods

Spiking simulations

All network simulations and models were implemented in the NEST simulator [96]. Neuronal dynamics are integrated by the Runge-Kutta-Fehlberg method as implemented in the GSL library [97] (gsl_odeiv_step_rkf45) – this forward integration scheme is used in the NEST simulator for all conductance-based models (at the time of writing). The short-term plasticity model is integrated exactly, based on inter-spike intervals. Code for network simulations will be made available upon publication at `https://github.com/EPFL-LCN/pub-seeholzer2018`.

Simulation protocol In all experiments (except those involving bi-stability, see below) spiking networks were simulated for a transient initial period of $t_{\text{initial}} = 500ms$. To center the network in an attractor state at a given angle $-\pi \leq \varphi < \pi$, we gave an initial cue signal by stimulating $0.1 \cdot N_E$ neurons centered at φ by strong excitatory input mediated by additional Poisson firing onto AMPA receptors ($1s$, 2kHz) with connections scaled down by a factor of $g_{\text{signal}} = 0.5$. The external input ceased at $t = t_{\text{off}} = 1.5s$. For simulations to estimate the diffusion we simulated until $t_{\max} = 15s$, yielding $13.5s$ of delay activity after the cue offset. For simulations to estimate drift we set $t_{\max} = 8s$, yielding $6.5s$ of delay activity after the cue offset.

For simulations exploring the bi-stability between the uniform state and a bump state (Fig. 1B1), we added an additional input prior to the spontaneous state. We stimulated simultaneously 20 excitatory neurons around 4 equally spaced cue points each (80 neurons in total, 500ms, 1.5kHz, AMPA connections scaled by a factor $g_{\text{signal}} = 2$). This was applied to settle networks into the uniform state more stably – without this perturbation, networks sometimes approached the bump state after being uniformly initialized. In both figures, we show population activity only after this initial stimulus was applied.

Estimation of centers and mean bump shapes To estimate centers of bump states, simulations were run until $t = t_{\max}$ and spikes were recorded from the excitatory population and converted to firing rates by convolving them with an exponential kernel ($\tau = 100ms$) [98] and then sampled at resolution 1ms. This results in vectors of firing rates $\nu_j(t), 0 \leq j \leq N_E - 1$ for every time t. We calculated the population center $\varphi(t)$ for time t by measuring the phase of the first spatial Fourier coefficient of the firing rates. This is given by $\varphi(t) = \arg\left(\sum_j \exp(i\frac{2\pi}{N_E}j)\nu_j(t)\right) - \pi$. For all analyses below, we identify $t = 0$ to be the time $t = t_{\text{off}}$ of the initial cue.

To measure the mean bump shapes, we first rectified the vectors $\nu_j(t)$ for every t by rotating the vector until $\varphi(t) = 0$. We then sampled the rectified firing rates starting from $1s$ after cue offset at intervals of 20ms, which were used to calculate the mean firing rates. S1 Fig shows mean rates for each simulation averaged over the ~ 1000 repetitions performed in the diffusion estimation (below).

Exclusion of bump trajectories Sometimes bump trajectories would leave the attractor state and return to the uniform state. We identified these trajectories in all experiments by identifying maximal firing rates across the population that dropped below 10Hz during the delay period. The such identified repetitions were excluded from the analyses, which occurred mostly in networks with no facilitation for $\tau_x = 150ms, \tau_u = 650ms$: at $U = 1$, we excluded 222/1000 repetitions from the diffusion estimation, while for all other $U \leq 0.8$ at most 15/1000 were excluded. Increasing the depression time constant also lead to less stable attractor states: for $\tau_x = 200ms, \tau_u = 650ms$ and $U = 0.8$, we had to exclude 250/1000 repetitions. During the simulations for drift estimation, we observed that frozen noise also leads to less stable bumps under weak facilitation for random and sparse connectivity ($p \ll 1$) and high leak variability ($\sigma_L \gg 0$).

Diffusion estimation Diffusion was estimated for each combination of network parameters by simulating 1000 repetitions (10 initial cue positions, 100 repetitions each) of $13.5s$ of delay activity. Center positions $\varphi_k(t)$ were estimated for each repetition k as described above. We then calculated for each repetition the offset relative to the position at $500ms$ by $\Delta\varphi_k(t) = \varphi_k(t - 500ms) - \varphi_k(500ms)$, effectively discarding the first $500ms$ after cue-offset. The time-dependent variance of K repetitions (excluding those repetitions in which the bump state was lost, see above) was then calculated as

$V(t) = \frac{1}{K}\sum_k \Delta\varphi_k^2(t)$. The diffusion strength can then be estimated from the slope of a linear least-squares regression (using the Scipy method *scipy.stats.linregress* [99]) to the variance as a function of time: $V(t) \approx D_0 + D \cdot t$, where the intercept D_0 is included to account for initial transients. We estimated confidence intervals by bootstrapping [100]: sampling K elements out of the K repetitions with replacement (5000 samples) and estimating the confidence level of 0.95 by the bias corrected and accelerated bootstrap implemented in *scikits-bootstrap* [101]. As a control, we calculated confidence intervals for D additionally by Jackknifing: after building a distribution of estimates of D on K one-left-out samples of all repetitions, the standard error of the mean can be calculated and is multiplied by 1.96 to obtain the 95% confidence interval [102] – confidence intervals obtained by this method were almost indistinguishable from confidence intervals obtained by bootstrapping.

Drift estimation Drift was estimated numerically for each combination of network and frozen noise parameters by simulating 400 repetitions (20 initial cue positions, 20 repetitions each) of $6.5s$ of delay activity. Centers positions $\varphi_k(t)$ were estimated for all K repetitions (excluding those repetitions in which the bump state was lost, see above) as explained above. We then computed displacements in time by computing a set of discrete differences

$$\Delta\varphi_k = \{(\varphi_k\left[t_0 + (j+1)dt\right] - \varphi\left[t_0 + j \cdot dt\right])/dt \mid \forall j \in \mathbb{N}_0 : t_0 + (j+1)dt \leq t_{\max}\},$$

where we chose $dt = 1.5s$ and $t_0 \in \{500ms, 700ms, 900ms, \ldots, 1900ms\}$. All differences are calculated with periodic boundary conditions on the circle $[-\pi, \pi)$, i.e. the maximal difference was π/dt. We then calculated a binned mean (100 bins on the ring, unless mentioned otherwise) of differences calculated for all K trajectories, to approximate the drift-fields as a function of positions on the ring.

Mutual information measure

We are estimating the mutual information between a set of initial positions $x \in [0, 2\pi)$ and associated final positions $y(x) \in [0, 2\pi)$ of the trajectories of a continuous attractor network over a fixed delay period of T. For our results, we take $T = 6.5s$. We constructed binned and normalized histograms (with bin size $n = 100$, but see below) as approximate probability distributions of initial positions $p_i = p([i-1]\,\frac{2\pi}{n} \leq x < i\frac{2\pi}{n})$ and all final positions $q_i = p([i-1]\,\frac{2\pi}{n} \leq y < i\frac{2\pi}{n})$ (with bins indexed by $1 \leq i \leq n$), as well as the bivariate probability distribution $r_{ij} = p([i-1]\,\frac{2\pi}{n} \leq x < i\frac{2\pi}{n}, [j-1]\,\frac{2\pi}{n} \leq y(x) < j\frac{2\pi}{n})$.

Using these, we can calculate the mutual information as [54, 55] $\mathrm{MI} = \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}\log_2\left(\frac{r_{ij}}{p_i q_j}\right)$. Note, that the sum effectively counts only nonzero entries of r_{ij} (trajectories that started in bin i and ended in bin j): these imply that $p_i \neq 0$ (a trajectory started in bin i) and $q_j \neq 0$ (a trajectory ended in bin j), which makes the sum well defined. Although the value of MI depends on the number of bins n, in Fig. 5 and Fig. 6 we normalize MI to that of the reference network ($U = 1$, no frozen noise, see *Short-term plasticity controls memory retention*), which leaves the resulting plot nearly invariant under a change of bin numbers.

Numerical integration of Langevin equations

Numerically integration of the homogeneous Langevin equations (Eq. (4)) describing drift and diffusion of bump positions $\varphi \in [-\pi, \pi)$ (with circular boundary conditions) has been implemented as a C extension in Cython [103] to the Python language [104]. Since the drift fields $A(\varphi)$ are estimated on a discretization of the interval $[-\pi, \pi)$ into

N bins, we first interpolate drift fields A given as N discretized values to obtain continuous fields – interpolations are obtained using cubic splines on periodic boundary conditions using the class *gsl_interp_cspline_periodic* of the Gnu Scientific Library [97].

For forward integration of the Langevin equation Eq. (4) from time $t = 0$, we start from an initial position $\varphi_0 = \varphi(t = 0)$. Given a time resolution dt (unless otherwise stated we use $dt = 0.1s$) and a maximal time $t_{\max}$ we repeat the following operations until we reach $t = t_{\max}$:

$$\begin{aligned} t &\to t + dt, \\ \varphi &\to \varphi + dt \cdot A(\varphi) + \sqrt{dtB} \cdot r, \\ \varphi &\to ((\varphi + \pi) \mod 2\pi) - \pi. \end{aligned}$$

Here, for each iteration r is a random number drawn from a normal distribution with zero mean and unit variance ($\langle r \rangle = 0$ and $\langle r^2 \rangle = 1$). The last step is performed to implement the circular boundary conditions on $[-\pi, \pi)$.

Code implementing this numerical integration scheme will be made available upon publication at `https://github.com/EPFL-LCN/pub-seeholzer2018-langevin`.

Distractor analysis

For the distractor analysis in Fig. 7, we let 40 neurons centered at the distractor position $\varphi_D = \frac{360°}{N} j - 180°$ fire at rates increased by 20Hz, yielding a vector of firing rate perturbations $\Delta\phi_{0,i} = 20\text{Hz}$ if $|i - j| \leq 20$ and $\Delta\phi_{0,i} = 0\text{Hz}$ otherwise. The vectors $\Delta\phi_{0,i}$ foreach distractor position φ_D are then used in Eq. (7) to calculate the corresponding drift fields. To calculate the final position φ_1 after 250ms of presenting the distractor, we generate 1000 trajectories starting from $\varphi_0 = 0$ by integrating the Langevin equation Eq. (4) for 250ms ($dt = 0.01$), the final positions of which are used to measure mean and standard deviation of φ_1. For the broader bump in Fig. 7D, we stretched (and interpolated) the firing rates ϕ_0 as well as the associated vectors J_0 and ϕ_0' along the x-axis to obtain vectors for bumps of the desired width, and then re-calculated the values of $\frac{dJ_0}{d\varphi}$.

Acknowledgments

The authors thank Tilo Schwalger and Johanni Brea for helpful discussions and feedback. Research was supported by the European Union Seventh Framework Program (FP7) under grant agreement no. 604102 (Human Brain Project, M.D.) and by the Swiss National Science Foundation (200020_147200, A.S.).

References

1. Goldman-Rakic PS. Cellular Basis of Working Memory. Neuron. 1995;14(3):477–485. doi:10.1016/0896-6273(95)90304-6.

2. Constantinidis C, Wang XJ. A Neural Circuit Basis for Spatial Working Memory. The Neuroscientist. 2004;10(6):553–65. doi:10.1177/1073858404268742.

3. Chaudhuri R, Fiete I. Computational Principles of Memory. Nature Neuroscience. 2016;19(3):394–403. doi:10.1038/nn.4237.

4. Curtis CE, D'Esposito M. Persistent Activity in the Prefrontal Cortex during Working Memory. Trends in Cognitive Sciences. 2003;7(9):415–423. doi:10.1016/S1364-6613(03)00197-9.

5. Durstewitz D, Seamans JK, Sejnowski TJ. Neurocomputational Models of Working Memory. Nature Neuroscience. 2000;3:1184–91. doi:10.1038/81460.

6. Barak O, Tsodyks M. Working Models of Working Memory. Current Opinion in Neurobiology. 2014;25:20–24. doi:10.1016/j.conb.2013.10.008.

7. Duarte R, Seeholzer A, Zilles K, Morrison A. Synaptic Patterning and the Timescales of Cortical Dynamics. Current Opinion in Neurobiology. 2017;43:156–165. doi:10.1016/j.conb.2017.02.007.

8. Amari Si. Dynamics of Pattern Formation in Lateral-Inhibition Type Neural Fields. Biological cybernetics. 1977;87:77–87.

9. Camperi M, Wang XJ. A Model of Visuospatial Working Memory in Prefrontal Cortex: Recurrent Network and Cellular Bistability. Journal of Computational Neuroscience. 1998;5(4):383–405.

10. Hansel D, Sompolinsky H. Modeling Feature Selectivity in Local Cortical Circuits. In: Koch C, Segev I, editors. Methods in Neural Modeling. from Synapses to Networks. MIT Press; 1998. p. 499–567.

11. Compte A, Brunel N, Goldman-Rakic PS, Wang XJ. Synaptic Mechanisms and Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model. Cerebral Cortex. 2000;10:910–923.

12. Samsonovich A, McNaughton BL. Path Integration and Cognitive Mapping in a Continuous Attractor Neural Network Model. Journal of Neuroscience. 1997;17(15):5900–20.

13. Stringer SM, Trappenberg TP, Rolls ET, de Araujo IET. Self-Organizing Continuous Attractor Networks and Path Integration: One-Dimensional Models of Head Direction Cells. Network. 2002;13(2):217–42.

14. Burak Y, Fiete IR. Accurate Path Integration in Continuous Attractor Network Models of Grid Cells. PLOS Computational Biology. 2009;5(2):e1000291. doi:10.1371/journal.pcbi.1000291.

15. Ben-Yishai R, Bar-Or RL, Sompolinsky H. Theory of Orientation Tuning in Visual Cortex. Proceedings of the National Academy of Sciences. 1995;92(9):3844–8.

16. Zhang K. Representation of Spatial Orientation by the Intrinsic Dynamics of the Head-Direction Cell Ensemble: A Theory. Journal of Neuroscience. 1996;16(6):2112–2126.

17. Seung HS. Continuous Attractors and Oculomotor Control. Neural Networks. 1998;11(7):1253–1258. doi:10.1016/S0893-6080(98)00064-1.

18. Knierim JJ, Zhang K. Attractor Dynamics of Spatially Correlated Neural Activity in the Limbic System. Annual Review of Neuroscience. 2012;35:267–85. doi:10.1146/annurev-neuro-062111-150351.

19. Moser EI, Roudi Y, Witter MP, Kentros C, Bonhoeffer T, Moser MB. Grid Cells and Cortical Representation. Nature Reviews Neuroscience. 2014;15(7):466–481. doi:10.1038/nrn3766.

20. Burak Y. Spatial Coding and Attractor Dynamics of Grid Cells in the Entorhinal Cortex. Current Opinion in Neurobiology. 2014;25:169–175. doi:10.1016/j.conb.2014.01.013.

21. Wu S, Wong KYM, Fung CCA, Mi Y, Zhang W. Continuous Attractor Neural Networks: Candidate of a Canonical Model for Neural Information Representation. F1000Research. 2016;5. doi:10.12688/f1000research.7387.1.

22. Wimmer K, Nykamp DQ, Constantinidis C, Compte A. Bump Attractor Dynamics in Prefrontal Cortex Explains Behavioral Precision in Spatial Working Memory. Nature Neuroscience. 2014;17(3):431–439. doi:10.1038/nn.3645.

23. Yoon K, Buice MA, Barry C, Hayman R, Burgess N, Fiete IR. Specific Evidence of Low-Dimensional Continuous Attractor Dynamics in Grid Cells. Nature Neuroscience. 2013;16(8):1077–1084. doi:10.1038/nn.3450.

24. Seelig JD, Jayaraman V. Neural Dynamics for Landmark Orientation and Angular Path Integration. Nature. 2015;521(7551):186–191. doi:10.1038/nature14446.

25. Kim SS, Rouault H, Druckmann S, Jayaraman V. Ring Attractor Dynamics in the Drosophila Central Brain. Science. 2017; p. eaal4835. doi:10.1126/science.aal4835.

26. Macoveanu J, Klingberg T, Tegnér J. A Biophysical Model of Multiple-Item Working Memory: A Computational and Neuroimaging Study. Neuroscience. 2006;141(3):1611–8. doi:10.1016/j.neuroscience.2006.04.080.

27. Wei Z, Wang XJ, Wang DH. From Distributed Resources to Limited Slots in Multiple-Item Working Memory: A Spiking Network Model with Normalization. Journal of Neuroscience. 2012;32(33):11228–40. doi:10.1523/JNEUROSCI.0735-12.2012.

28. Roggeman C, Klingberg T, Feenstra HEM, Compte A, Almeida R. Trade-off between Capacity and Precision in Visuospatial Working Memory. Journal of Cognitive Neuroscience. 2014;26(2):211–222. doi:10.1162/jocna00485.

29. Almeida R, Barbosa Ja, Compte A. Neural Circuit Basis of Visuo-Spatial Working Memory Precision: A Computational and Behavioral Study. Journal of Neurophysiology. 2015;114(3):1806–1818. doi:10.1152/jn.00362.2015.

30. Cano-Colino M, Almeida R, Compte A. Serotonergic Modulation of Spatial Working Memory: Predictions from a Computational Network Model. Frontiers in Integrative Neuroscience. 2013;7. doi:10.3389/fnint.2013.00071.

31. Cano-Colino M, Almeida R, Gomez-Cabrero D, Artigas F, Compte A. Serotonin Regulates Performance Nonmonotonically in a Spatial Working Memory Network. Cerebral Cortex. 2014;24(9):2449–2463. doi:10.1093/cercor/bht096.

32. Murray JD, Anticevic A, Gancsos M, Ichinose M, Corlett PR, Krystal JH, et al. Linking Microcircuit Dysfunction to Cognitive Impairment: Effects of Disinhibition Associated with Schizophrenia in a Cortical Working Memory Model. Cerebral Cortex. 2012;doi:10.1093/cercor/bhs370.

33. Cano-Colino M, Compte A. A Computational Model for Spatial Working Memory Deficits in Schizophrenia. Pharmacopsychiatry. 2012;45(S 01):S49–S56.

34. Tsodyks M, Sejnowski T. Associative Memory and Hippocampal Place Cells. Neural Systems. 1995;6:81–86.

35. Spiridon M, Gerstner W. Effect of Lateral Connections on the Accuracy of the Population Code for a Network of Spiking Neurons. Network. 2001;12(4):409–21.

36. Renart A, Song P, Wang XJ. Robust Spatial Working Memory through Homeostatic Synaptic Scaling in Heterogeneous Cortical Networks. Neuron. 2003;38(3):473–85.

37. Wu S, Hamaguchi K, Amari Si. Dynamics and Computation of Continuous Attractors. Neural Computation. 2008;20(4):994–1025.

38. Itskov V, Hansel D, Tsodyks M. Short-Term Facilitation May Stabilize Parametric Working Memory Trace. Frontiers in Computational Neuroscience. 2011;5(October):40–40. doi:10.3389/fncom.2011.00040.

39. Burak Y, Fiete IR. Fundamental Limits on Persistent Activity in Networks of Noisy Neurons. Proceedings of the National Academy of Sciences. 2012;109(43):17645–17650.

40. Kilpatrick ZP, Ermentrout B. Wandering Bumps in Stochastic Neural Fields. arXiv:12053072 [math, nlin, q-bio]. 2012;.

41. Brody C, Romo R, Kepecs A. Basic Mechanisms for Graded Persistent Activity: Discrete Attractors, Continuous Attractors, and Dynamic Representations. Current Opinion in Neurobiology. 2003; p. 204–211. doi:10.1016/S0959-4388(03)00050-3.

42. York LC, van Rossum MCW. Recurrent Networks with Short Term Synaptic Depression. Journal of Computational Neuroscience. 2009;27(3):607–620. doi:10.1007/s10827-009-0172-4.

43. Romani S, Tsodyks M. Short-Term Plasticity Based Network Model of Place Cells Dynamics. Hippocampus. 2015;25(1):94–105. doi:10.1002/hipo.22355.

44. Barbieri F, Brunel N. Irregular Persistent Activity Induced by Synaptic Excitatory Feedback. Frontiers in Computational Neuroscience. 2007;1. doi:10.3389/neuro.10.005.2007.

45. Hansel D, Mato G. Short-Term Plasticity Explains Irregular Persistent Activity in Working Memory Tasks. The Journal of Neuroscience. 2013;33(1):133–49. doi:10.1523/JNEUROSCI.3455-12.2013.

46. Pereira J, Wang XJ. A Tradeoff Between Accuracy and Flexibility in a Working Memory Circuit Endowed with Slow Feedback Mechanisms. Cerebral Cortex. 2015;25(10):3586. doi:10.1093/cercor/bhu202.

47. Gerstner W, Kistler WM. Spiking Neuron Models: Single Neurons, Populations, Plasticity. Cambridge University Press; 2002.

48. Tsodyks M, Pawelzik K, Markram H. Neural Networks with Dynamic Synapses. Neural Computation. 1998;10(4):821–835. doi:10.1162/089976698300017502.

49. Brunel N, Wang X. Effects of Neuromodulation in a Cortical Network Model of Object Working Memory Dominated by Recurrent Inhibition. Journal of Computational Neuroscience. 2001;11:63–85.

50. Richardson M. Firing-Rate Response of Linear and Nonlinear Integrate-and-Fire Neurons to Modulated Current-Based and Conductance-Based Synaptic Drive. Physical Review E. 2007;76(2):1–15. doi:10.1103/PhysRevE.76.021919.

51. Wang XJ. Synaptic Basis of Cortical Persistent Activity: The Importance of NMDA Receptors to Working Memory. The Journal of Neuroscience. 1999;19(21):9587–603.

52. van Kampen NG. Stochastic Processes in Physics and Chemistry. 2nd ed. North Holland; 1992.

53. Gardiner C. Stochastic Methods: A Handbook for the Natural and Social Sciences. 4th ed. Springer; 2009.

54. Latham PE, Roudi Y. Mutual Information. Scholarpedia. 2009;4(1):1658.

55. Cover TM, Thomas JA. Elements of Information Theory. John Wiley & Sons; 2012.

56. Kilpatrick ZP, Ermentrout B, Doiron B. Optimizing Working Memory with Heterogeneity of Recurrent Cortical Excitation. Journal of Neuroscience. 2013;33(48):18999–19011. doi:10.1523/JNEUROSCI.1641-13.2013.

57. Macoveanu J, Klingberg T, Tegnér J. Neuronal Firing Rates Account for Distractor Effects on Mnemonic Accuracy in a Visuo-Spatial Working Memory Task. Biological Cybernetics. 2007;96(4):407–419. doi:10.1007/s00422-006-0139-8.

58. Ploner CJ, Gaymard B, Rivaud S, Agid Y, Pierrot-Deseilligny C. Temporal Limits of Spatial Working Memory in Humans. European Journal of Neuroscience. 1998;10(2):794–797. doi:10.1046/j.1460-9568.1998.00101.x.

59. White JM, Sparks DL, Stanford TR. Saccades to Remembered Target Locations: An Analysis of Systematic and Variable Errors. Vision Research. 1994;34(1):79–92. doi:10.1016/0042-6989(94)90259-3.

60. Wang Y, Markram H, Goodman PH, Berger TK, Ma J, Goldman-Rakic PS. Heterogeneity in the Pyramidal Network of the Medial Prefrontal Cortex. Nature Neuroscience. 2006;9(4):534–42. doi:10.1038/nn1670.

61. Yang CR, Seamans JK, Gorelova N. Electrophysiological and Morphological Properties of Layers V-VI Principal Pyramidal Cells in Rat Prefrontal Cortex in Vitro. Journal of Neuroscience. 1996;16(5):1904–1921.

62. Dégenètais E, Thierry AM, Glowinski J, Gioanni Y. Electrophysiological Properties of Pyramidal Neurons in the Rat Prefrontal Cortex: An In Vivo Intracellular Recording Study. Cerebral Cortex. 2002;12(1):1–16. doi:10.1093/cercor/12.1.1.

63. Collins CE, Turner EC, Sawyer EK, Reed JL, Young NA, Flaherty DK, et al. Cortical Cell and Neuron Density Estimates in One Chimpanzee Hemisphere. Proceedings of the National Academy of Sciences. 2016;113(3):740–745. doi:10.1073/pnas.1524208113.

64. Funahashi S, Bruce CJ, Goldman-Rakic PS. Mnemonic Coding of Visual Space in the Monkey's Dorsolateral Prefrontal Cortex. Journal of Neurophysiology. 1989; p. 331–349.

65. Chafee MV, Goldman-Rakic PS. Matching Patterns of Activity in Primate Prefrontal Area 8a and Parietal Area 7ip Neurons During a Spatial Working MemoryTask. Journal of Neurophysiology. 1998;79(6):2919–2940.

66. Wibisono A, Jog V, Loh PL. Information and Estimation in Fokker-Planck Channels. arXiv preprint arXiv:170203656. 2017;.

67. Sandamirskaya Y. Dynamic Neural Fields as a Step toward Cognitive Neuromorphic Architectures. Frontiers in Neuroscience. 2014;7. doi:10.3389/fnins.2013.00276.

68. Pan B, Zucker RS. A General Model of Synaptic Transmission and Short-Term Plasticity. Neuron. 2009;62(4):539–554. doi:10.1016/j.neuron.2009.03.025.

69. Brunton BW, Botvinick MM, Brody CD. Rats and Humans Can Optimally Accumulate Evidence for Decision-Making. Science. 2013;340(6128):95–98. doi:10.1126/science.1233912.

70. Selemon LD, Goldman-Rakic PS. Common Cortical and Subcortical Targets of the Dorsolateral Prefrontal and Posterior Parietal Cortices in the Rhesus Monkey: Evidence for a Distributed Neural Network Subserving Spatially Guided Behavior. Journal of Neuroscience. 1988;8(11):4049–4068.

71. Roach JP, Ben-Jacob E, Sander LM, Zochowski MR. Formation and Dynamics of Waves in a Cortical Model of Cholinergic Modulation. PLoS Comput Biol. 2015;11(8):e1004449. doi:10.1371/journal.pcbi.1004449.

72. Lim S, Goldman MS. Balanced Cortical Microcircuitry for Spatial Working Memory Based on Corrective Feedback Control. Journal of Neuroscience. 2014;34(20):6790–6806. doi:10.1523/JNEUROSCI.4602-13.2014.

73. Anwar H, Li X, Bucher D, Nadim F. Functional Roles of Short-Term Synaptic Plasticity with an Emphasis on Inhibition. Current Opinion in Neurobiology. 2017;43:71–78. doi:10.1016/j.conb.2017.01.002.

74. Nadim F, Bucher D. Neuromodulation of Neurons and Synapses. Current Opinion in Neurobiology. 2014;29:48–56. doi:10.1016/j.conb.2014.05.003.

75. Oh M, Zhao S, Matveev V, Nadim F. Neuromodulatory Changes in Short-Term Synaptic Dynamics May Be Mediated by Two Distinct Mechanisms of Presynaptic Calcium Entry. Journal of Computational Neuroscience. 2012;33(3). doi:10.1007/s10827-012-0402-z.

76. Kreitzer AC, Regehr WG. Modulation of Transmission during Trains at a Cerebellar Synapse. Journal of Neuroscience. 2000;20(4):1348–1357.

77. Brenowitz S, David J, Trussell L. Enhancement of Synaptic Efficacy by Presynaptic GABAB Receptors. Neuron. 1998;20(1):135–141. doi:10.1016/S0896-6273(00)80441-9.

78. Barrière G, Tartas M, Cazalets JR, Bertrand SS. Interplay between Neuromodulator-Induced Switching of Short-Term Plasticity at Sensorimotor Synapses in the Neonatal Rat Spinal Cord. The Journal of Physiology. 2008;586(Pt 7):1903–1920. doi:10.1113/jphysiol.2008.150706.

79. Hempel CM, Hartman KH, Wang XJ, Turrigiano GG, Nelson SB. Multiple Forms of Short-Term Plasticity at Excitatory Synapses in Rat Medial Prefrontal Cortex. Journal of Neurophysiology. 2000;83(5):3031–3041.

80. Sakurai A, Katz PS. State-, Timing-, and Pattern-Dependent Neuromodulation of Synaptic Strength by a Serotonergic Interneuron. Journal of Neuroscience. 2009;29(1):268–279. doi:10.1523/JNEUROSCI.4456-08.2009.

81. Laing CR, Longtin A. Noise-Induced Stabilization of Bumps in Systems with Long-Range Spatial Coupling. Physica D: Nonlinear Phenomena. 2001;160(3):149–172.

82. Wang H, Lam K, Fung CCA, Wong KYM, Wu S. Rich Spectrum of Neural Field Dynamics in the Presence of Short-Term Synaptic Depression. Physical Review E. 2015;92(3):032908. doi:10.1103/PhysRevE.92.032908.

83. Mi Y, Lin X, Wu S. Neural Computations in a Dynamical System with Multiple Time Scales. Frontiers in Computational Neuroscience. 2016;10. doi:10.3389/fncom.2016.00096.

84. Fung C, Wong K, Wu S. A Moving Bump in a Continuous Manifold: A Comprehensive Study of the Tracking Dynamics of Continuous Attractor Neural Networks. Neural Computation. 2010;22(3):752–792.

85. Brunel N. Dynamics of Sparsely Connected Networks of Excitatory and Inhibitory Spiking Neurons. Journal of Computational Neuroscience. 2000;8(3):183–208.

86. Wang XJ. Synaptic Reverberation Underlying Mnemonic Persistent Activity. Trends in Neurosciences. 2001;24(8):455–63.

87. Renart A, Moreno-Bote R, Wang XJ, Parga N. Mean-Driven and Fluctuation-Driven Persistent Activity in Recurrent Networks. Neural Computation. 2007;19(1):1–46. doi:10.1162/neco.2007.19.1.1.

88. Barbieri F, Brunel N. Can Attractor Network Models Account for the Statistics of Firing during Persistent Activity in Prefrontal Cortex? Frontiers in neuroscience. 2008;2(1):114–122.

89. Compte A, Constantinidis C, Tegnér J, Raghavachari S, Chafee MV, Goldman-Rakic PS, et al. Temporally Irregular Mnemonic Persistent Activity in Prefrontal Neurons of Monkeys During a Delayed Response Task. Journal of Neurophysiology. 2003;90(5):3441–3454. doi:10.1152/jn.00949.2002.

90. Mongillo G, Hansel D, van Vreeswijk C. Bistability and Spatiotemporal Irregularity in Neuronal Networks with Nonlinear Synaptic Transmission. Physical Review Letters. 2012;108(15):158101. doi:10.1103/PhysRevLett.108.158101.

91. Markram H, Wang Y, Tsodyks M. Differential Signaling via the Same Axon of Neocortical Pyramidal Neurons. Proceedings of the National Academy of Sciences. 1998;95(9):5323–8.

92. Mongillo G, Barak O, Tsodyks M. Synaptic Theory of Working Memory. Science. 2008;319(5869):1543–1546. doi:10.1126/science.1150769.

93. Gerstner W, Kistler WM, Naud R, Paninski L. Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition. Cambridge University Press; 2014.

94. Seeholzer A, Deger M, Gerstner W. Efficient Low-Dimensional Approximation of Continuous Attractor Networks. arXiv:171108032 [q-bio]. 2017;.

95. Forster O. Analysis 1: Differential- und Integralrechnung einer Veränderlichen. Springer-Verlag; 2016.

96. Bos H, Morrison A, Peyser A, Hahne J, Helias M, Kunkel S, et al.. NEST 2.10.0; 2015. https://doi.org/10.5281/zenodo.44222.

97. Galassi M, Davies J, Theiler J, Gough B, Jungman G, Booth M, et al. GNU Scientific Library Reference Manual. 3rd ed.; 2009.

98. Nawrot M, Aertsen A, Rotter S. Single-Trial Estimation of Neuronal Firing Rates: From Single-Neuron Spike Trains to Population Activity. Journal of Neuroscience Methods. 1999;94:81–92.

99. Jones E, Oliphant T, Peterson. SciPy.Org — SciPy.Org; 2017. http://scipy.org/.

100. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. CRC Press; 1994.

101. Evans C. Scikits-Bootstrap; 2017. https://github.com/cgevans/scikits-bootstrap.

102. Abdi H, Williams L. Jackknife. Encyclopedia of research design. 2010; p. 1–10.

103. Behnel S, Bradshaw R, Citro C, Dalcin L, Seljebotn DS, Smith K. Cython: The Best of Both Worlds. Computing in Science Engineering. 2011;13(2):31–39. doi:10.1109/MCSE.2010.118.

104. Oliphant TE. Python for Scientific Computing. Computing in Science & Engineering. 2007;9(3):10–20. doi:10.1109/MCSE.2007.58.

Supporting information

S1 Fig. Spiking networks produce similar stable firing rate profiles across parameters. For each choice of short-term plasticity parameters U, τ_u, and τ_x, we tuned the recurrent conductances (g_{EE}, g_{EI}, g_{IE}, g_{II}) and the width σ_w of the distance-dependent weights (cf. Eq. (32)) such that the "bump" shape of the stable firing rate profile is close to a generalized Gaussian $\nu(\theta) = g_0 + g_1 \exp\left(-\left[\frac{|\theta|}{g_\sigma}\right]^{g_r}\right)$ with parameters $g_0 = 0.1\mathrm{Hz}$, $g_1 = 40.0\mathrm{Hz}$, $g_\sigma = 0.5$, $g_r = 2.5$. See *Optimization of network parameters* in Materials and Methods for details, S2 Table for parameter values after tuning, and S1 Table for parameters that stay constant. **A** After tuning, the resulting firing rate profiles for different parameter values of U and τ_u are very similar. Averaged mean firing rates in bump state, measured from ~ 1000 spiking simulations. **A1-A3** Remaining slight parameter-dependent changes of bump shapes, measured by fitting the generalized Gaussian $\nu(\theta)$ to the measured firing rate profiles displayed in A. **A1** Top firing rate g_1. **A2** Half-width parameter g_σ. **A3** Sharpness parameter g_r. **B and B1-B3** Same as in A and A1-A3, for additional variation of the depression time scale τ_x.

S2 Fig. Theoretical prediction of diffusion strength as a function of STP parameters. All color values display diffusion magnitude estimated from B in Eq. (4) with bump shape estimated from the reference network ($U = 1, \tau_x = 150\mathrm{ms}$, compare 3B,C, dashed lines). Units of color values are $\frac{\mathrm{idx}^2}{s}$ with values of level lines as indicated. **A** Diffusion as function of facilitation U and depression time constant τ_x. Facilitation time constant was $\tau_u = 650ms$. **B** Diffusion as function of facilitation U and facilitation time constant τ_u. Depression time constant was $\tau_x = 150ms$. **C** Diffusion as function of depression time constant τ_x and facilitation time constant τ_u. Facilitation U was $U = 0.5$.

S3 Fig. Comparison of theoretically predicted fields to simulations. A Averaged root mean square error (RMSE) between predicted fields (Eq. (7)) and fields extracted from simulations (mean over 18-20 networks, error bars show 95% confidence of the mean). Both frozen noise parameters (σ_L and $1-p$) are plotted on the same x-axis. **B** Normalized RMSE: each RMSE is normalized by the range (max − min) of the joint data of simulated and predicted fields it is calculated on. Colors as in A. **C** Average RMSE (same data as in A) plotted as a function of the mean expected field magnitude (estimated separately for each network, then averaged). Colors as in A. **D** Worst (top) and best (bottom) match between predicted field (blue line) and field extracted from simulations (black line) of the group with the largest mean RMSE in panels A, C ($U = 1, 1-p = 0.75$). Shaded areas show 1 standard deviation of points included in the binned mean estimate (100 bins) of the extracted field.

S4 Fig. Mutual information normalized to compare slopes. Same data as in Fig. 5B, but MI is normalized to the average MI of each spiking network without heterogeneities (leftmost dot for each green, orange, and blue group of curves/dots), making explicitly visible the change in slope of the drop-off as heterogeneity parameters are increased. Dashed lines connect the means, for visual guidance.

S5 Fig. Theoretical predictions of working memory stability. All panels show theoretically predicted expected displacement over 1 second (Eq. (11)) for networks with random and sparse connections ($p = 0.12$) and leak reversal potential heterogeneity ($\sigma_L = 1.7mV$). White lines show displacement contour lines for $1, 2$ and $5\,\text{deg}$. **A** Displacement as a function of the facilitation time constant τ_u and facilitation U for $\tau_x = 150ms$ and $N = 5000$.**B** Displacement as a function of system size and facilitation U for $\tau_x = 150ms$ and $\tau_u = 650ms$. **C-D** Displacement as a function of depression time constant τ_x and facilitation U for $N = 5000$ (C) and $N = 20000$ (D). In both panels $\tau_u = 650ms$.

S6 Fig. Dependence of diffusion strength B on shape parameters. Diffusion was calculated from Eq. (5) with bump solutions $\phi_0 = g_1 \exp(-\left|\frac{x}{g_\sigma}\right|^{g_r})$. The values of $\frac{dJ_0}{d\varphi}$ and ϕ_0' were calculated by fitting and extrapolating (linearly, for $\phi_0 > 40.31$Hz) curves $\phi_0 \to \phi_0'$ and $\phi_0 \to J_0$ that were obtained from the numerical values extracted for $g_1 = 40.31$Hz, $g_\sigma = 0.51$ by theory (see *Firing rate approximation* in Materials and Methods). Thus, any nonlinearity or saturation of the inputs and input-output relation for $\phi_0 > 40.31$Hz was not included. This approximate analysis shows that the major dependence of the diffusion expected in the system is on the bump width g_σ, although a minor dependence on g_1 is seen.

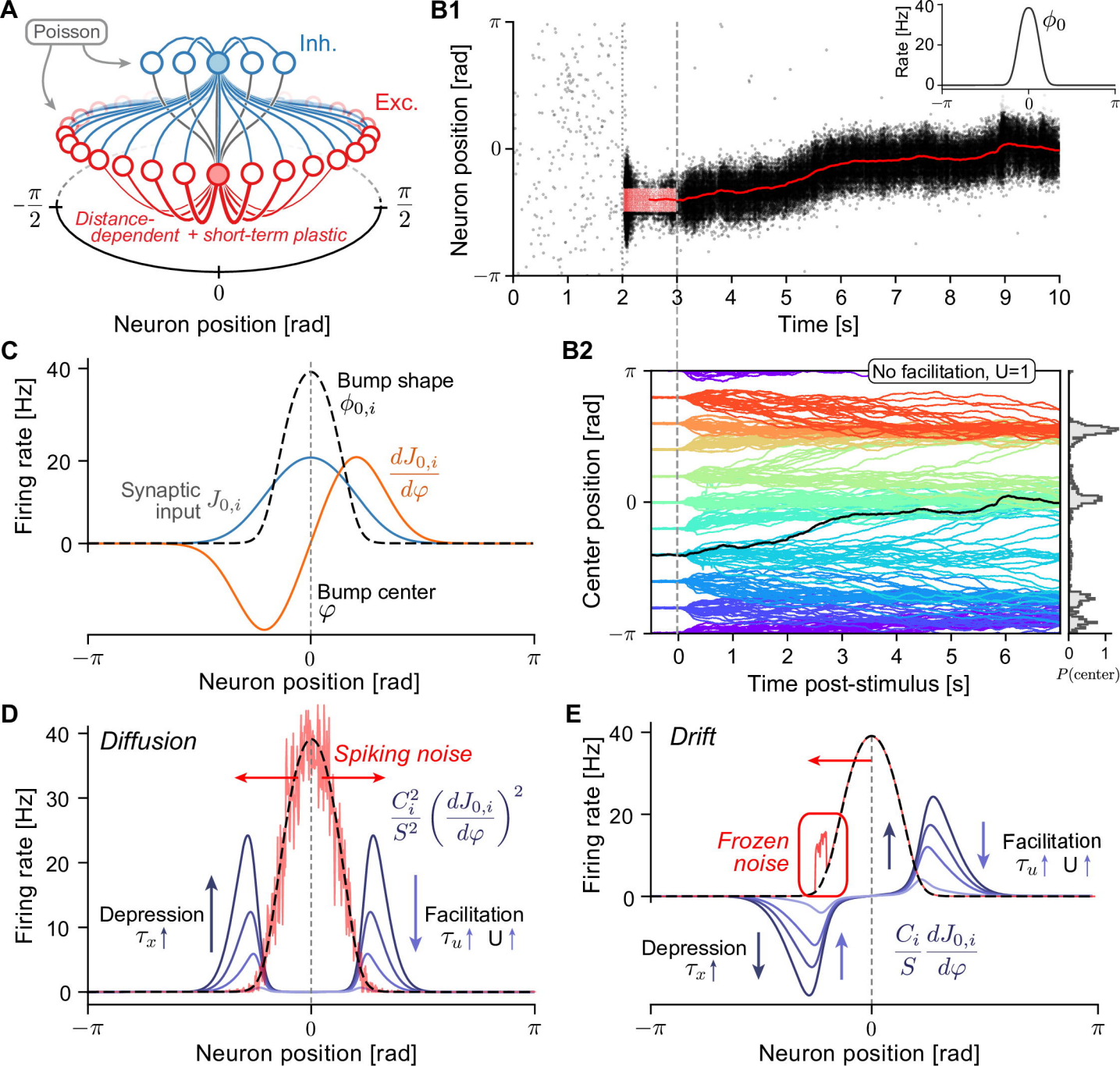
S7 Fig. Short-term plasticity does not affect spiking statistics. Mean firing rate, coefficient of variation of the inter-spike interval distribution (CV), and local CV (CV_2 [89]) for two attractor networks with different STP parameters. All measures were computed on spike-trains measured over a period of $4s$, recorded $500ms$ after offset of the external input which was centered at angle 0. Across STP parameters, networks display similarly reduced CVs for increased mean firing rates, leading to large CVs for neurons located in the flanks of the firing rate profile and low CVs for neurons located near the center. **A** Networks with large diffusion coefficient ($U = 0.8$, $\tau_u = 650ms$, $\tau_x = 200ms$) that underwent non-stationary diffusion during the recording of spikes: the measured mean firing rates (gray line) differ visibly from the firing rates estimated after centering the firing rate distribution at each point in time. Due to this non-stationarity, CVs at intermediate firing rates appear elevated, while the

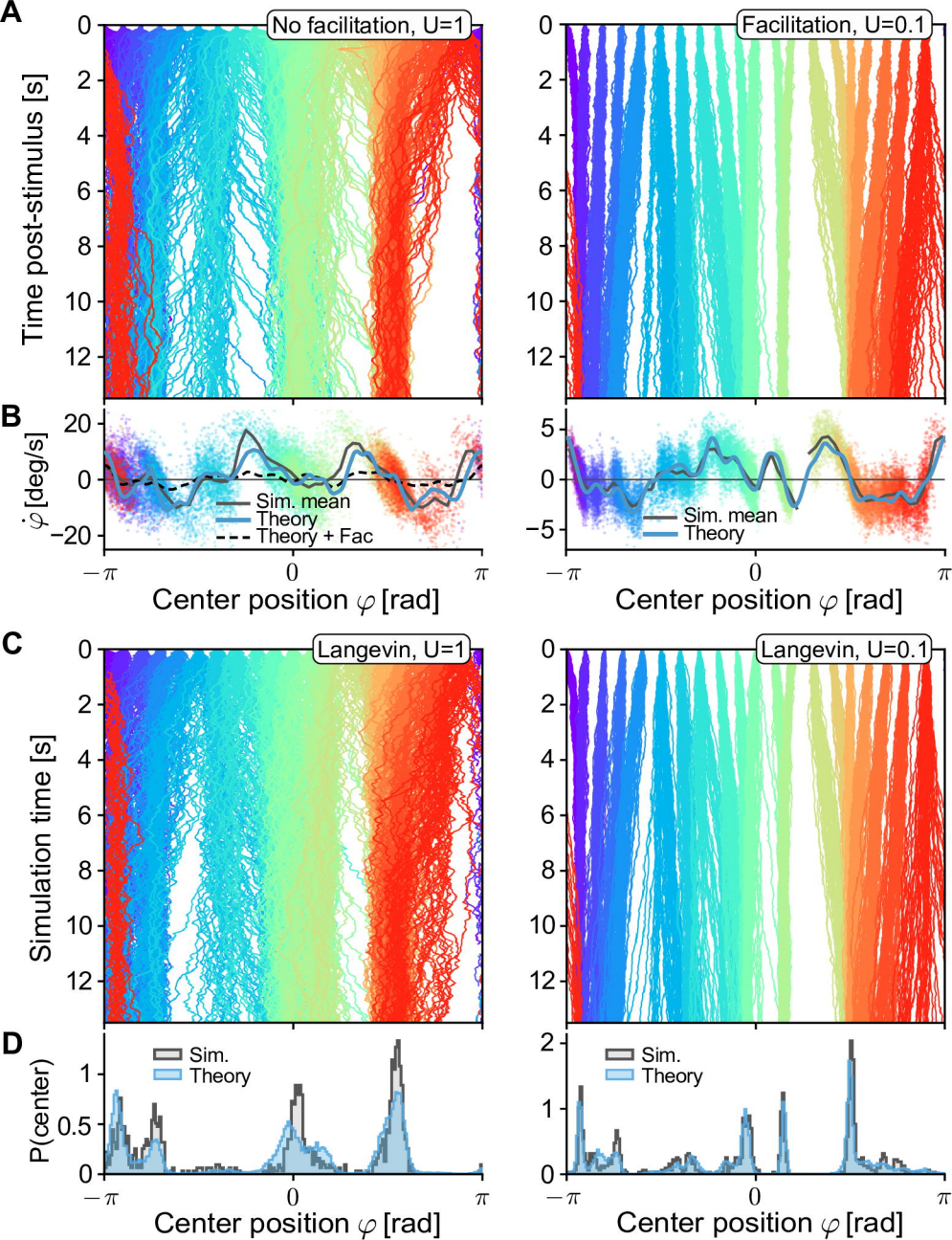
local CV (CV_2) shows values close to stationary networks (see B). **B** The same network as in A, with strong facilitation ($U = 0.1$). Reduced diffusion leads to a nearly stationary firing rate profile, and coincident CV and CV_2 measures.

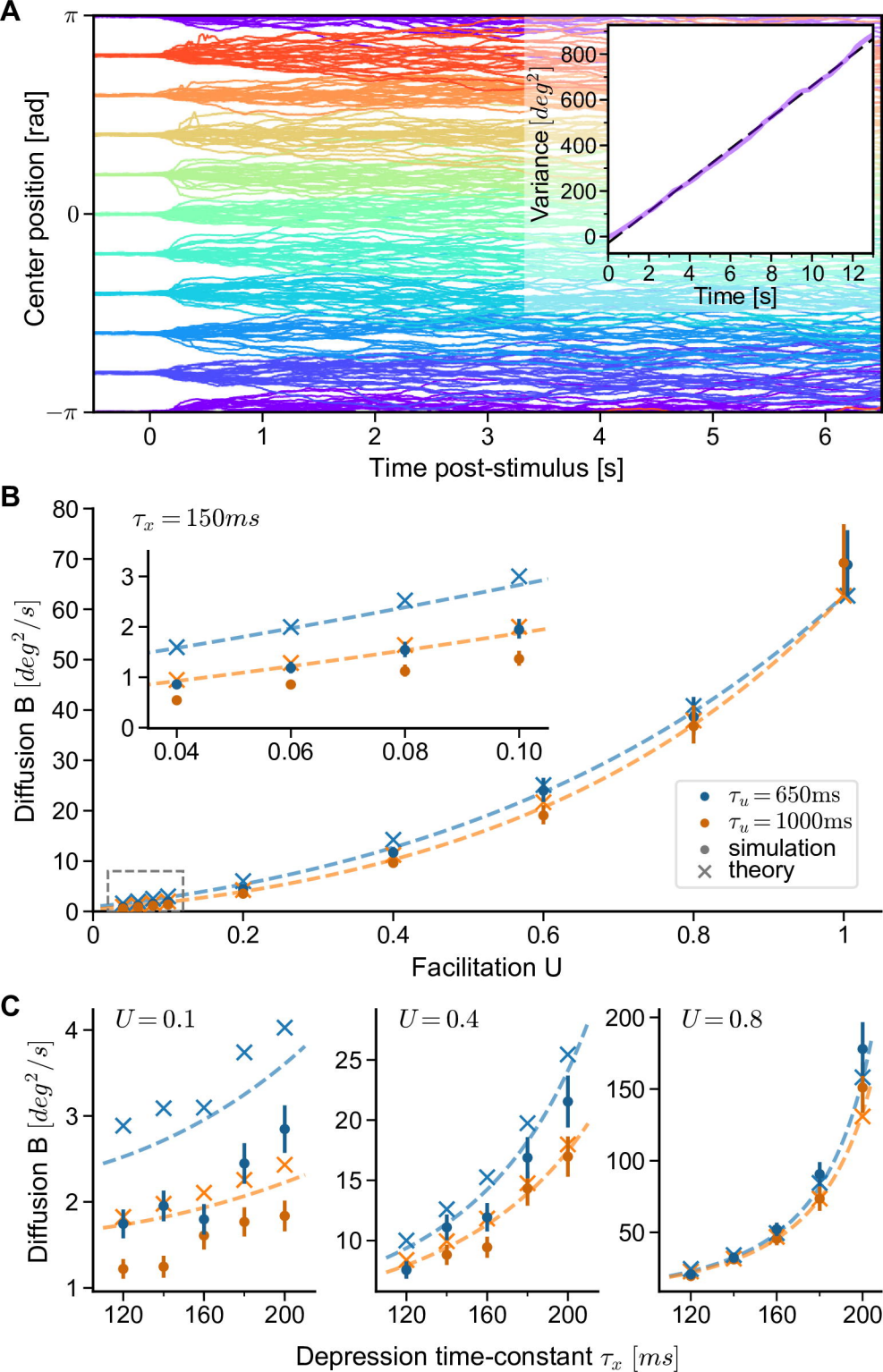
S1 Supporting Information. Detailed mathematical derivations.

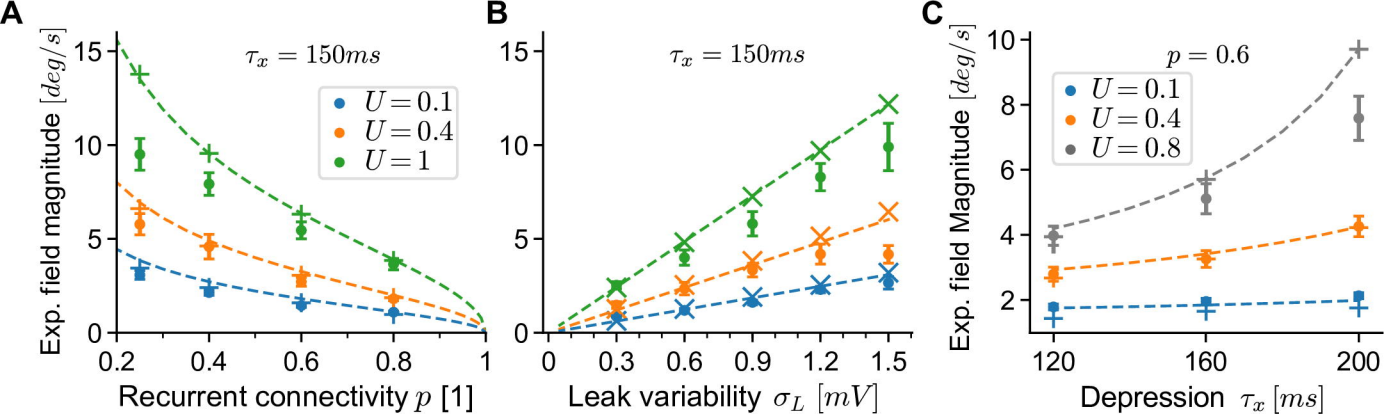
S1 Table. Parameters for spiking simulations. Parameter values are modified from [11] and [49]. For recurrent conductances see the table in S2 Table.

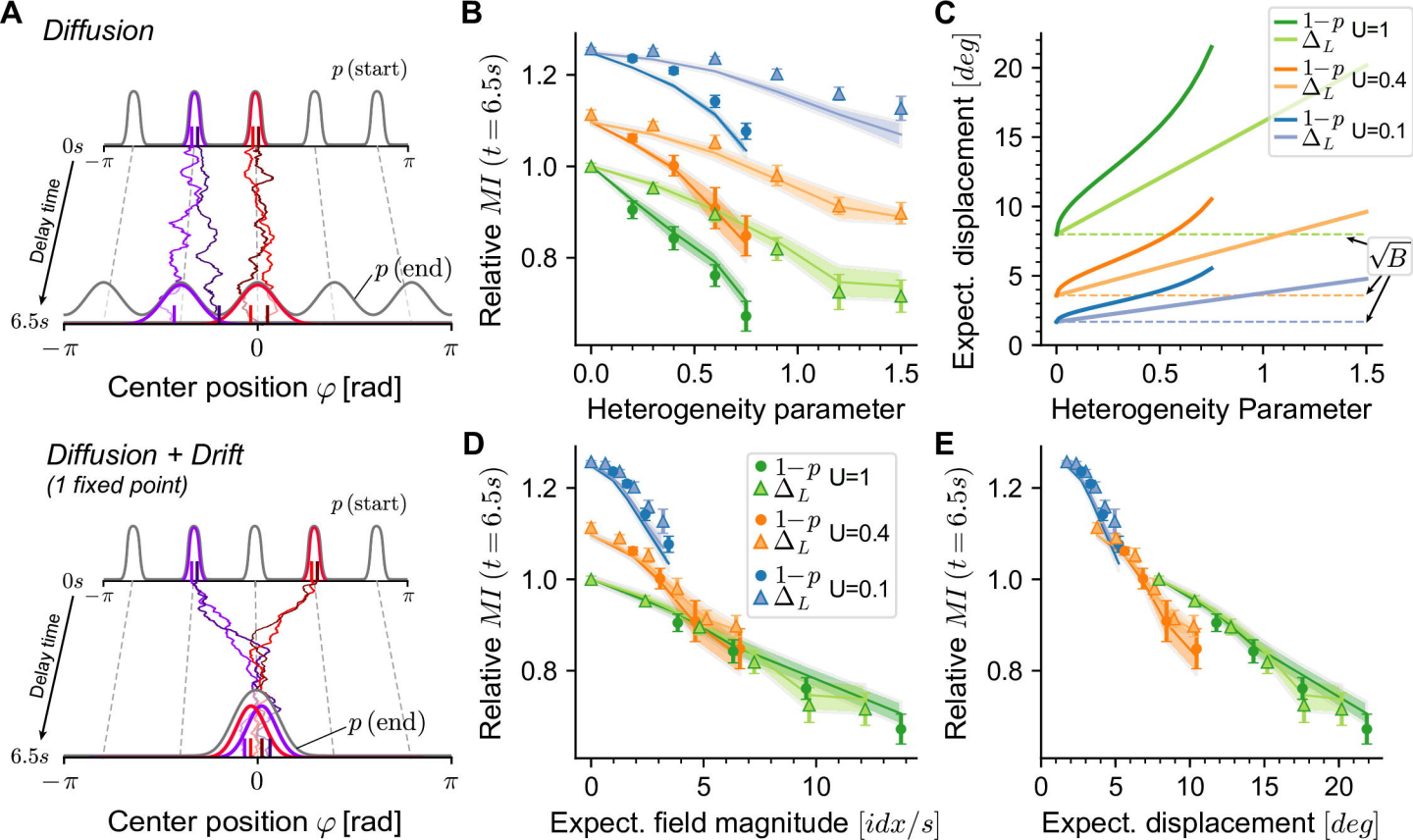
S2 Table. Conductance and connectivity parameters for spiking simulations. For all networks we set $w_+ = 4.0$. Recurrent conductance parameters are given for combinations of short-term plasticity parameters according to the following notation. g_{EE}: excitatory conductance g_E on excitatory neurons; g_{IE}: excitatory conductance g_E on inhibitory neurons; g_{EI}: inhibitory conductance g_I on excitatory neurons; g_{II}: inhibitory conductance g_I on inhibitory neurons.

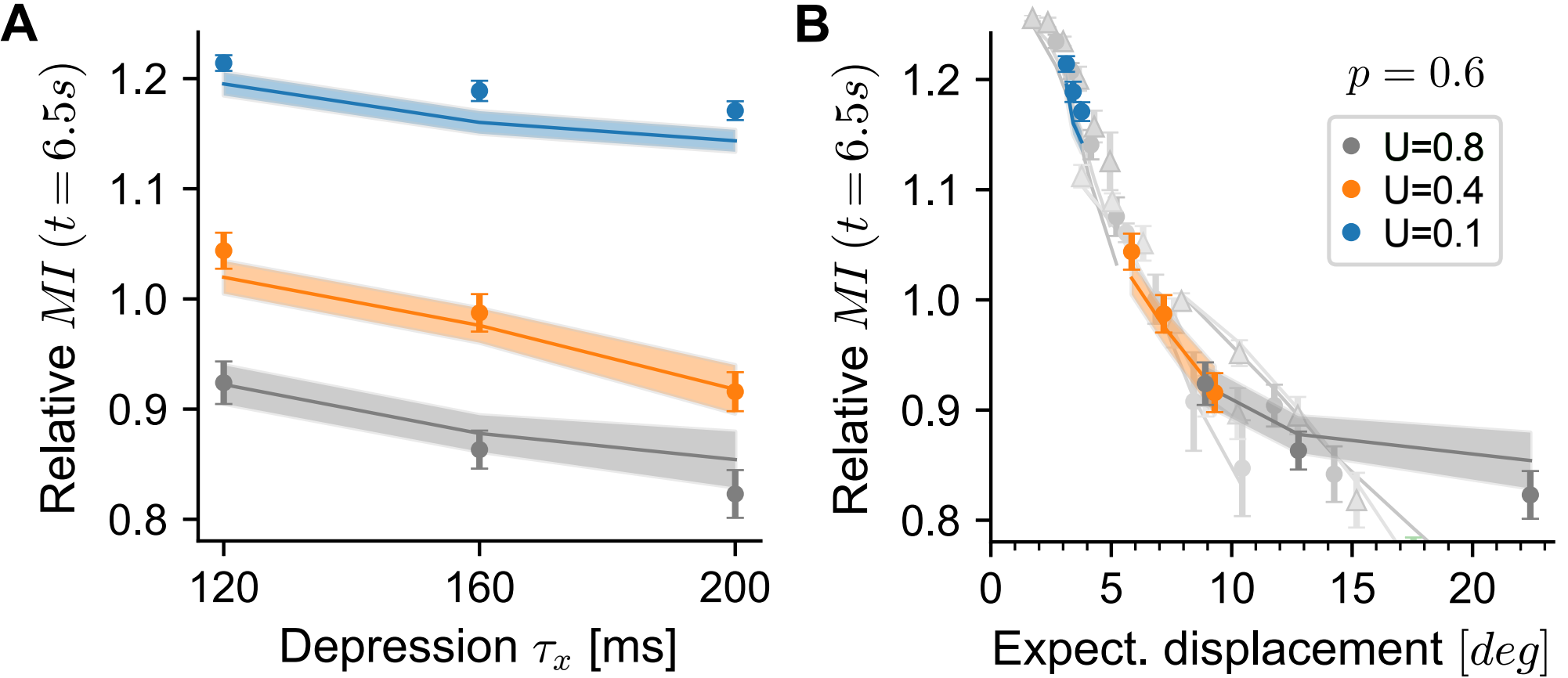

A
Poisson
Inh.
Exc.
Distance-dependent + short-term plastic
Neuron position [rad]
B1
Neuron position [rad]
Time [s]
Rate [Hz]
ϕ_0
B2
No facilitation, U=1
Center position [rad]
Time post-stimulus [s]
P(center)
C
Firing rate [Hz]
Bump shape
$\phi_{0,i}$
Synaptic input $J_{0,i}$
$\frac{dJ_{0,i}}{d\varphi}$
Bump center
φ
Neuron position [rad]
D
Diffusion
Spiking noise
$\frac{C_i^2}{S^2}\left(\frac{dJ_{0,i}}{d\varphi}\right)^2$
Depression
$\tau_x\uparrow$
Facilitation
$\tau_u\uparrow$ U$\uparrow$
Firing rate [Hz]
Neuron position [rad]
E
Drift
Frozen noise
Facilitation
$\tau_u\uparrow$ U$\uparrow$
Depression
$\tau_x\uparrow$
$\frac{C_i}{S}\frac{dJ_{0,i}}{d\varphi}$
Firing rate [Hz]
Neuron position [rad]

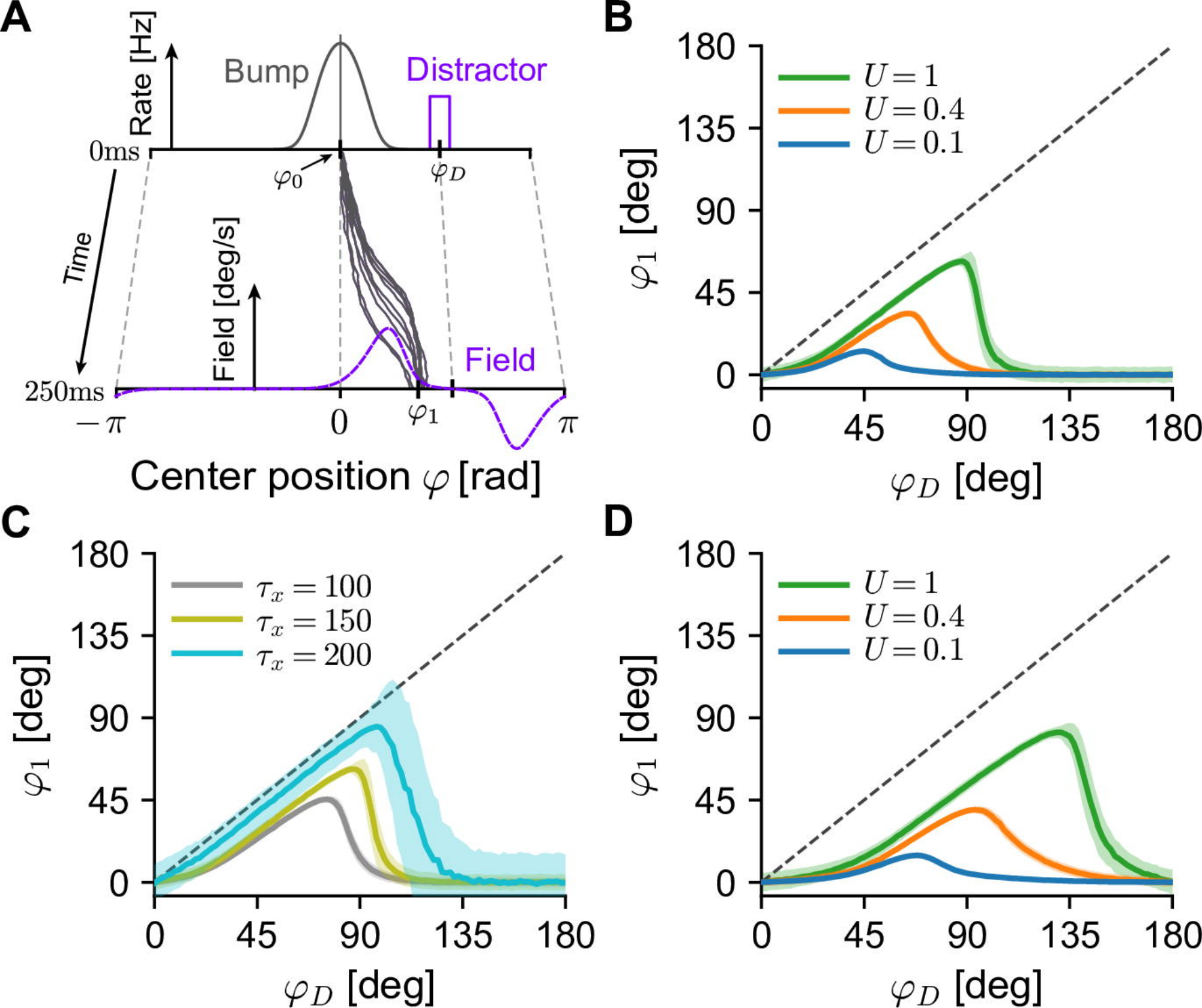

A
No facilitation, U=1
Facilitation, U=0.1
Time post-stimulus [s]
B
$\dot{\varphi}$ [deg/s]
Sim. mean
Theory
Theory + Fac
Center position φ [rad]
C
Langevin, U=1
Langevin, U=0.1
Simulation time [s]
D
P(center)
Sim.
Theory
Center position φ [rad]


A
Center position [rad]
Time post-stimulus [s]
Variance [deg^2]
Time [s]
B
$\tau_x = 150ms$
Diffusion B [deg^2/s]
Facilitation U
$\tau_u = 650$ms
$\tau_u = 1000$ms
simulation
theory
C
$U = 0.1$
$U = 0.4$
$U = 0.8$
Diffusion B [deg^2/s]
Depression time-constant τ_x [ms]


A
B
C
$\tau_x = 150ms$
$\tau_x = 150ms$
$p = 0.6$
$U = 0.1$
$U = 0.4$
$U = 1$
$U = 0.1$
$U = 0.4$
$U = 0.8$
Exp. field magnitude $[deg/s]$
Exp. field Magnitude $[deg/s]$
Recurrent connectivity p [1]
Leak variability σ_L $[mV]$
Depression τ_x $[ms]$
0
5
10
15
2
4
6
8
10
0.2
0.4
0.6
0.8
1
0
0.3
0.6
0.9
1.2
1.5
120
160
200


A
Diffusion
p (start)
0s
−π
π
Delay time
p (end)
6.5s
−π
0
π
Center position φ [rad]
Diffusion + Drift
(1 fixed point)
p (start)
0s
−π
π
Delay time
p (end)
6.5s
−π
0
π
Center position φ [rad]
B
Relative MI (t = 6.5s)
1.2
1.0
0.8
0.0
0.5
1.0
1.5
Heterogeneity parameter
C
Expect. displacement [deg]
20
15
10
5
0
1−p
Δ_L
U=1
1−p
Δ_L
U=0.4
1−p
Δ_L
U=0.1
√B
0
0.5
1
1.5
Heterogeneity Parameter
D
Relative MI (t = 6.5s)
1.2
1.0
0.8
1−p
Δ_L
U=1
1−p
Δ_L
U=0.4
1−p
Δ_L
U=0.1
0
5
10
Expect. field magnitude [idx/s]
E
Relative MI (t = 6.5s)
1.2
1.0
0.8
0
5
10
15
20
Expect. displacement [deg]


A
B
Relative MI $(t = 6.5s)$
1.2
1.1
1.0
0.9
0.8
120
160
200
Depression τ_x [ms]
$p = 0.6$
U=0.8
U=0.4
U=0.1
0
5
10
15
20
Expect. displacement $[deg]$


A
Rate [Hz]
Bump
Distractor
0ms
φ_0
φ_D
Time
Field [deg/s]
Field
250ms
$-\pi$
0
φ_1
π
Center position φ [rad]
B
$U = 1$
$U = 0.4$
$U = 0.1$
φ_1 [deg]
φ_D [deg]
C
$\tau_x = 100$
$\tau_x = 150$
$\tau_x = 200$
φ_1 [deg]
φ_D [deg]
D
$U = 1$
$U = 0.4$
$U = 0.1$
φ_1 [deg]
φ_D [deg]