Robust maintenance of both stimulus location and amplitude in a working memory model based on dendritic bistability

Jiacheng Xu,[1*] Daniel L Cox,[1] Steven J Luck,[2] Mark S Goldman,[3,4*]

[1]Department of Physics and Astronomy,
University of California Davis, Davis CA 95616 USA
[2]Center for Mind & Brain and Department of Psychology,
University of California Davis, Davis CA 95616 USA
[3]Center for Neuroscience, and Department of Neurobiology, Physiology, and Behavior,
University of California, Davis, Davis, CA 95616, USA
[4]Department of Ophthalmology and Vision Science,
University of California Davis, Davis CA 95616 USA

*To whom correspondence should be addressed:
E-mail: jiacheng.xu@dartmouth.edu, msgoldman@ucdavis.edu

Abstract

Working memory is a core feature of cognition that enables items to be maintained and manipulated over short durations of time. Stored information can be binary, such as the presence or absence of an object, or graded, such as the graded intensity or location of a feature. Current computational models of working memory cannot robustly maintain both the graded intensity and spatial location of a stored item. Here, we show how this limitation can be overcome if neurons contain multiple bistable dendritic compartments. First, we illustrate the core mechanism for the storage of graded amplitude information in a simple spiking "autapse" circuit consisting of a single neuron connected to itself. Second, we reduce this model to a rate-based model that permits analytic understanding. Third, we implement this mechanism within a spatially extended architecture in which the spatial location of an item is encoded by the set of active neurons. In contrast to classic spatial working memory models, which only encode the bi-

nary presence of an item at a given location, the multi-dendrite-neuron model robustly encodes both the amplitude and location of an item in working memory in a noise-resistant manner and without requiring fine tuning of parameters. We show analytically that the key mechanism permitting the storage of amplitude information is equivalent to that of the simpler autapse circuit. This work provides a solution to the problem of encoding graded information in spatial working memory and demonstrates how dendritic computation can increase the representational capacity and robustness of working memory.

Significance Statement

Animals can readily remember both the location of an item and analog features of the item such as its amplitude or intensity. Remarkably, current computational models of working memory require extreme fine tuning of model parameters to perform this task. Here, we show how this limitation can be overcome if neurons have active dendritic processes that enable local dendritic compartments of the neuron to robustly maintain digital "up" ("plateau potential") or "down" states. By activating a variable number of plateau potentials at each location, the models can jointly maintain both the location and amplitude of a stimulus in working memory in a noise-resistant manner. This work demonstrates how local dendritic processes can enhance the computational capabilities of working memory networks.

Introduction

Working memory refers to a mechanism that provides a temporary storage buffer for cognitive tasks such as search, reasoning, and decision-making. Working memory often must precisely store analog quantities such as the location, amplitude, or color of an object. The mechanisms by which such analog quantities are stored robustly in working memory has been puzzling, as the continuous attractor mechanisms traditionally used to describe the dynamics of analog working memory require fine tuning to avoid instability and are subject to drift in the presence of noise [1, 2, 3, 4, 5].

Classic models of graded working memory are based on line attractors. In a line attractor, stored memories are encoded into a continuous set of stably maintainable neural activity states, where each state represents the value of a distinct analog feature. Such models typically fall into two classes. In one class, which we term "location-coding" [2, 6], the encoded quantity is represented by a fixed amplitude pattern of neural activity that can shift between different neural locations, with the location of the pattern representing the analog stimulus [7]. This class of models has commonly been used to describe the encoding of variables such as the spatial location of an object [8], the orientation of an animal's head in space [9], or the location (hue) of a color on a color wheel [10]. The canonical model of this class is the "ring" model in which neurons are arranged in a one-dimensional ring, with each neuron preferentially encoding a particular location in a circular stimulus space (Fig. 1A). This model has found direct experimental support in the microscopic connectivity of the fly head direction system [11] and has been widely used to explain many cognitive tasks [12].

The second class of models, which we term "amplitude-encoding" [2, 3, 13], encodes memorized quantities in the amplitude of a largely fixed spatial pattern of neural activity. This class of models has been used to represent the memory of amplitude- or intensity-like variables such as luminance [14] or vibration frequency [15]. Such models have direct anatomical support in the connectome of the oculomotor integrator of the eye movement system [16] and are the canonical neural implementation of accumulation-of-evidence based decision-making [17, 18].

In actual organisms, working memory can simultaneously encode both the location and amplitude of an item. Further, even when encoding only the location of an item, Bayesian theories

posit that the amplitude of the spatial pattern of neural activity represents the precision with which the location is encoded [19, 20]. Despite this, few models exist of such joint amplitude-location encoding, and the few existing models exhibit a lack of robustness in maintaining analog amplitudes of neural firing. This lack of robustness reflects fundamental limitations of classic models of location and amplitude encoding. Classic location-coding models like the ring model exhibit only a single all-or-none "bump" of activity and thus do not represent amplitude information (Fig. 1B). Further, they require either a perfectly symmetric network connectivity [7] or a functionally similar connectivity [4, 21, 22] to maintain a graded representation of location. Classic amplitude-coding models require fine tuning of their recurrent synaptic connectivity to achieve stable encoding of analog stimulus amplitudes – recurrent feedback that is too weak or too strong leads to rapid decay or growth, respectively, of neural activity on the timescale of the intrinsic neuronal or synaptic time constants of the model. Previous models of joint amplitude-location encoding have inherited these shortcomings and either require fine tuning of parameters to store analog amplitudes [23, 24, 25, 26] or trade off sensitivity in representation against robustness [27]. Further, even when fine-tuned, such models are subject to rapid diffusion of representations in the presence of noisy neural activity.

A separate set of amplitude-coding memory models has succeeded in overcoming the fine tuning problem by incorporating bistable memory components. In these models, analog memory values are generated through summing up the contributions of many bistable neurons [13] or dendritic compartments [28, 29]. These models inherit the robustness of the bistable processes upon which they are built, while also being capable of storing a graded representation of amplitude limited only by the number of bistable components. However, in location-coding memory models, only cellular bistability has been employed [30, 31], which stabilizes the memory of a given location, but does not permit the joint memorization of stimulus amplitude. Biophysically, bistability can be generated by positive feedback among small clusters of neurons [13] or, at a cellular level, through positive-feedback interactions mediated by the ionic conductances that underlie somatic [32, 33, 34, 35, 36, 37, 38, 39, 40] or local dendritic [41, 42, 43] plateau potentials.

Here, we show how a network of neurons with bistable dendrites provides, to our knowl-

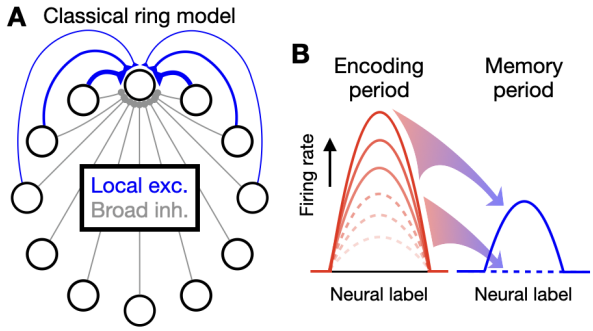

Figure 1: Schematic overview of inability of classic ring models to memorize graded amplitudes. **A**, Classical ring model. Neurons are arranged in a one-dimensional ring with local excitation between nearby neurons (blue) and broad inhibition across all neurons (gray). **B**, Inability to memorize graded amplitudes. In classical ring models, only two stable memory states exist: a fixed amplitude pattern of activity (solid blue line) that occurs when the encoding period neural activities, caused by different stimulus amplitudes, have high firing rates (solid red lines), or an untuned baseline level of activity (dashed blue line) that occurs when the encoding period activities have low firing rates (dashed red lines). Graded values of stimulus amplitudes are not memorized.

edge, the first model that jointly and robustly encodes location and amplitude information. We first demonstrate the core amplitude-storage biophysical mechanism through construction of a spiking, multi-dendrite ”autapse” model consisting of a single neuron connected to itself [1]. We then simplify this biophysical model to an analytically solvable firing rate model that provides rigorous insight into the mechanism of amplitude storage. We next construct a multi-neuron ring model and show how, through a mechanism that can be directly mapped onto that of the simple autapse model, the network can represent both the analog amplitude and analog location of a stimulus. The model exhibits robustness to noise in its encoding of both memory amplitude and location, thus demonstrating how bistable neuronal processes can enhance the performance of working memory.

Materials and Methods

Here we provide the equations and numerical methods underlying the three models described in the main text. First, we present the equations and simulation methods for the spiking autapse circuit model. Second, we present the equations governing the dynamics of the rate-based autapse circuit model, including those that describe the multistable band that geometrically determines the model's stability properties. Third, we present the dynamical equations for the multistable ring model. Finally, we present numerical simulation specifications for the models, including characterizing the ring model's behavior in the presence of noise. Mathematical derivations of the exact or approximate behaviors of the rate-based models are provided in the Mathematical Appendix.

Spiking autapse circuit model with multiple conductance-based, bistable dendrites

The spiking autapse circuit model consists of an integrate-and-fire soma with $N_d = 10$ conductance-based dendrites. The somatic voltage V_s is governed by

$$\begin{aligned} C\frac{dV_s}{dt} &= -I_L - I_d - I_{stim} - I_{ton} \\ &= -\bar{g}_L(V_s - E_L) - \sum_{i=1}^{N_d} \bar{g}_{ds}(V_s - V_{d,i}) - \bar{g}_{stim} s_{stim}(V_s - E_{AMPA}) - I_{ton}, \end{aligned} \tag{1}$$

where I_L is the leak current, I_d is the summed current received from all dendrites, I_{stim} is the excitatory current representing the external stimulus, which arrives through AMPA receptors, and I_{ton} is the tonic background current. We chose parameter values similar to those of a previous model [43]. The membrane capacitance is $C = 10$ nF/mm^2. The maximal conductances are $\bar{g}_L = 0.5$ μS/mm^2 and $\bar{g}_{stim} = 5.2$ μS/mm^2. The conductance mediating the flow of currents from dendrite to soma is $\bar{g}_{ds} = 0.007$ μS/mm^2. The equilibrium potentials are $E_L = -80$ mV and $E_{AMPA} = 0$ mV. The tonic current is $I_{ton} = -17.2$ nA/mm^2. The synaptic activation variable s_{stim} obeys dynamics described below. The firing threshold for the soma is $V_{th} = -50$ mV. Each spike sets the somatic voltage to 30 mV for 3 ms, after which it resets the somatic voltage to -80 mV.

The voltage dynamics of the ith dendritic compartment, $V_{d,i}$, were adapted from [43], with modified maximum conductance values and the inward rectifier potassium (Kir) conductance

simplified to not include a synaptically driven component:

$$C\frac{dV_{d,i}}{dt} = -I_{L,i} - I_{Kir,i} - I_{s,i} - I_{AMPA,i} - I_{NMDA,i}, \tag{2}$$

where the first two terms give the intrinsic dendritic leak current $I_{L,i}$ and inward-rectifying potassium current $I_{Kir,i}$, the third term $I_{s,i}$ gives the current arriving from the soma, and the final two terms give the AMPA and NMDA components of the synaptic currents driven recurrently by somatic spiking. These currents are described by the following equations:

$$\begin{aligned} I_{L,i} &= \bar{g}_L(V_{d,i} - E_L), \\ I_{Kir,i} &= \bar{g}_{Kir}\frac{V_{d,i} - E_{Kir}}{1 + e^{0.1(V_{d,i}+100)}}, \\ I_{s,i} &= \eta\bar{g}_{ds}(V_{d,i} - V_s), \\ I_{AMPA,i} &= \bar{g}_{AMPA}s_{AMPA}(V_{d,i} - E_{AMPA}), \\ I_{NMDA,i} &= \bar{g}_{NMDA,i}s_{NMDA}\frac{V_{d,i} - E_{NMDA}}{1 + 0.15e^{-0.08V_{d,i}}}, \end{aligned} \tag{3}$$

where $\eta = 2$ represents the ratio of the area of the soma to that of each dendrite. $E_{NMDA} =$ 0 mV, $E_{Kir} =$ –90 mV, $\bar{g}_{AMPA} = 0.38\ \mu S/mm^2$ and $\bar{g}_{Kir} = 8.6\ \mu S/mm^2$. The maximum conductance for the NMDA receptors is $\bar{g}_{NMDA,i} = \frac{48}{(i+4.5)^{0.54}}\ \mu S/mm^2$.

The synaptic activation s_α obeys:

$$\begin{aligned} \tau_\alpha\frac{ds_\alpha}{dt} &= -s_\alpha, \text{ at all times}, \\ \Delta s_\alpha &= 0.5(1 - s_\alpha), \text{ each time a spike is received}, \end{aligned} \tag{4}$$

where α denotes AMPA (s_{stim} and s_{AMPA}) or NMDA (s_{NMDA}) receptors with a decay time constant $\tau_{AMPA} = 2$ ms or $\tau_{NMDA} = 100$ ms, respectively.

To derive the individual dendrite bistable input-output relations shown in Figure 2B, we removed all of the self-connections shown in Figure 2A. Next, a periodic train of spikes whose firing frequency steps up and down across a wide range of firing frequencies was delivered to each dendrite while its dendritic voltage was recorded. During this process, the dendritic voltage is also affected by somatic voltage changes through the electrical coupling between the somatic and dendritic compartments. To assure that the dendritic input-output relation was assessed with realistic somatic voltage levels, we confirmed that the somatic spike train driven by the dendritic inputs resembled that of the periodic inputs delivered to the dendrites. For the simulation, V_s and $V_{d,i}$ were initialized to –70 mV, and the periodic train of input spikes delivered to each

dendrite started at a rate of 0 Hz. The spike train lasted for a period of 1400 ms, and the averaged dendritic voltages over the last 1000 ms were recorded for all dendrites simultaneously. Next, the frequency of the periodic input train was increased by 1 Hz, and the recording repeated. Once the frequency reached 70 Hz, it was decreased by 1 Hz at each recording until returning to 0 Hz.

To demonstrate that the circuit maintains multiple levels of activity (Fig. 2C), we implemented the following procedure: V_s and $V_{d,i}$ were initialized to –70 mV, and the circuit dynamics were allowed to stabilize for a period of 1200 ms before providing a sequence of external stimuli. Each external stimulus in the sequence consisted of a step on and off of periodic spikes for 400 ms, followed by an 800 ms memory period with no external stimulus. The frequency of the external stimulus sequence started at 20 Hz and increased by 5 Hz for each step until reaching 65 Hz. For the parameter values given above, this procedure activated one additional dendrite with each step increase in frequency, as shown in Figure 2C.

Rate-based autapse circuit model

SIMPLIFIED AUTAPSE DYNAMICS

The dendrites in the rate-based autapse circuit model were modeled with a simple bistable input-output relation that approximated that observed in the biophysical model. Specifically, each dendrite has a bistable input-output relation as a function of the firing rate r of the form (Fig. 3A):

$$D_i(r) = \begin{cases} \beta\Theta(w_i r - T_u), & \text{activation from down to up state,} \\ \beta\Theta(w_i r - T_d), & \text{deactivation from up to down state,} \end{cases} \tag{5}$$

where $\beta = 1$ is the dendritic contribution to firing rate for an up-state dendrite, $\Theta(x)$ indicates the Heaviside (unit step) function, w_i is the weight of the connection onto the ith dendrite, and $T_u = 9$ and $T_d = 2$ are constant up and down thresholds. Dendrites are ordered such that w_i is a monotonically decreasing function of the dendrite index i. A larger synaptic weight w_i, which corresponds to a larger value of $\bar{g}_{NMDA,i}$ in the spiking model, effectively lowers the up and down thresholds. We can then rewrite $D_i(r)$ in terms of threshold firing rates for activating or deactivating a dendrite,

$$D_i(r) = \begin{cases} \beta\Theta(r - r^{th}_{u,i}), & \text{activation from down to up state,} \\ \beta\Theta(r - r^{th}_{d,i}), & \text{deactivation from up to down state,} \end{cases} \tag{6}$$

where the effective up and down thresholds are:

$$r_{u,i}^{th} = \frac{T_u}{w_i} \text{ and } r_{d,i}^{th} = \frac{T_d}{w_i}. \tag{7}$$

For analytic simplicity, we assumed that the somatic firing rate dynamics are given by an exponential approach to a steady state given by the sum of the dendritic contributions and the input I_{stim} that represents the strength of the stimulus to be remembered by the circuit:

$$\tau \frac{dr}{dt} = -r + \sum_{i=1}^{N_d} D_i(r) + I_{stim}, \tag{8}$$

where N_d is the number of dendrites. τ is chosen to capture, in a simple manner, the slowest timescale of neural activity, which is set by the NMDA receptor kinetics in the spiking model. Explicit inclusion of additional synaptic or dendritic dynamics does not qualitatively change the primary results of this work.

DETERMINATION OF THE SHAPE OF THE MULTISTABLE BAND

The total dendritic feedback to the soma $D_{tot}(r) = \sum_{i=1}^{N_d} D_i(r)$ graphically defines a band of multistable memory storage. Here, we provide the methods by which the edges of this band and their continuum approximation (Fig. 3) are defined. Following Eq. (8), the right side of the band, which consists of the set of up thresholds of sequentially activated dendrites, is given by

$$B_u(r) = \sum_{i=1}^{N_d} \beta \Theta(r - r_{u,i}^{th}). \tag{9}$$

For a large number of dendrites, the weight function w_i and the edge of the band can be approximated by a continuous set of values, enabling the band to be described by a continuous functional form. Replacing the discrete index i by a continuum index $x \in [0, N_d]$ and assuming w_i changes gradually, we approximate the discrete set of w_i by a continuous monotonically decaying function $w(x)$. Then we have:

$$r_u^{th}(x) = \frac{T_u}{w(x)}, \tag{10}$$

and B_u becomes:

$$B_u = \int_0^{N_d} \beta \Theta[r - r_u^{th}(x)] dx. \tag{11}$$

To derive an analytical expression for the right side of the band, we evaluate the above integral. The integrand is zero for $r < r_u^{th}(0)$ and a constant value of βN_d for all values of x such that $r > r_u^{th}(N_d)$. For intermediate firing rate values, using Eq. (10), every dendrite with an x value satisfying $r > T_u/w(x)$ contributes to the integrand, so that the total number of activated dendrites equals $w^{-1}(T_u/r)$. Thus, the values of the right side of the multistable band, $B_u(r)$, correspond to β times the number of dendrites activated at a steady-state firing rate r:

$$B_u(r) = \begin{cases} 0, & \text{for } r < r_u^{th}(0), \\ \beta w^{-1}(\frac{T_u}{r}), & \text{for } r_u^{th}(0) \leq r \leq r_u^{th}(N_d), \\ \beta N_d, & \text{for } r > r_u^{th}(N_d). \end{cases} \tag{12}$$

Similar reasoning applies for deriving the left side of the band, $B_d(r)$.

Rate-based model with ring architecture

The ring architecture network underlying Figures 4-7 and illustrated in Figure 4A was constructed as follows. The network consists of N neurons, with each neuron having $N_d = N$ dendrites. As in traditional ring models, except in Figure 7C, the recurrent excitatory weights $w_{ij}(d)$ were chosen to be symmetric, with a falloff in strength as a function of the distance d between neurons along the periodically arranged ring. The recurrent inhibitory weights were chosen to be global and of uniform weight w_{inh}. Each excitatory connection projects to a separate bistable dendrite on the postsynaptic neuron, and each inhibitory connection as well as the external stimulus project to the soma of the postsynaptic neuron. The firing rate dynamics of each neuron is given by

$$\tau \frac{dr_i}{dt} = -r_i + \sum_{j=1}^{N} D_{ij}(r_j) - w_{inh} \sum_{j=1}^{N} r_j + I_{stim,i}, \tag{13}$$

where $D_{ij}(r_j)$ is the bistable dendritic input-output relation, and $I_{stim,i}$ is the external stimulus input. The dendritic relation $D_{ij}(r_j)$ for the contribution to the firing rate of neuron i from the dendrite receiving input from neuron j, is defined analogously to $D_i(r)$ in the rate-based autapse model:

$$D_{ij}(r_j) = \begin{cases} \beta\Theta(w_{ij}r_j - T_u), & \text{activation from down to up state,} \\ \beta\Theta(w_{ij}r_j - T_d), & \text{deactivation from up to down state.} \end{cases} \tag{14}$$

SIMULATION DETAILS IN THE NOISE-FREE CASE

For all ring network model simulations, we simulated $N = 360$ neurons, with $w_{inh} = 1/360$, $\beta = 1$, and $\tau = 50$ ms. The external stimulus to each neuron, $I_{stim,i}$, is a Gaussian function, $I_A e^{-\frac{(i-180)^2}{2\times 10^2}}$. The local excitatory weights are given by a power law function $w_{ij} = \frac{18}{(d_{ij}+2)^2}$, except for Figures 4F,G and 6. $T_u = 9$ and $T_d = 2$ as in the autapse model, except for Figure 7. The stimulus amplitude values simulated are $I_A = 8, 16, 24$, or 32 in Figure 4C and increase with a stepsize of 0.1 from 0 to 60 in Figure 4D.

The time course of each simulation was as follows: With all dendrites originally in the down state, $I_{stim,i}$ was applied for 1000 ms (the 'encoding period'). The stimulus was then turned off, and the subsequent 'memory period' activity was simulated for 1000 ms.

The approximate multistable band of network feedback shown in Figure 4B corresponds to the neuron $i = 180$ at the center of the bump of activity that determines the height of the activity bump. The B_u and B_d curves that define the edges of the band were obtained by simulation as follows:

- Tracing the up-threshold curve B_u. During the encoding period, we plotted the total recurrent feedback $\beta \sum_{j=1}^{N} \Theta[w_{ij} r_j - T_u] - w_{inh} \sum_{j=1}^{N} r_j$ in Eqs. (13) and (14) against the firing rate r_i, with time as the parametric variable. This numerically generated the up-threshold curve B_u. The stimulus amplitude used to generate this plot was $I_A = 80$.

- Tracing the down-threshold curve B_d. The firing rate was first allowed to reach its stable, steady-state value of memory activity. To trace B_d, we then applied a uniform inhibition, $I_{stim,i} = -10$ for 1000 ms, until the network activity returned to zero.

As in the autapse case, the amplitude of the decay term, r_i in Eq. (13), is represented by a line of unity slope.

In Figure 5A, we used stimulus amplitudes of $I_A = 20$, 40, and 80. In Figure 5B, we applied two successive stimuli centered at the same location. First, a smaller stimulus $I_A = 20$ was applied, leading the subsequent memory activity to stabilize at the location of the red dot. Then a larger stimulus with $I_A = 80$ was applied, followed by a second memory period.

For the analytical results in Figure 4, the external stimulus was a Gaussian with profile $R(x) = e^{-\frac{(x-180)^2}{2\times 10^2}}$, which has an integrated area $\kappa \approx 10\sqrt{2\pi}$.

In Figure 4F,G, simulations were based upon the weight function analytically calculated to produce the linear memory amplitude versus stimulus amplitude relationship $M_A = c(I_A - I_{A,th})$, with $c = 0.5$ and $I_{A,th} = 1.2$ (see Mathematical Appendix, Eq. (51)). In Figure 4G, I_A takes integer values from 0 to 35.

SIMULATIONS OF NETWORKS WITH NOISE

For all simulations with noise, spatially and temporally independent Gaussian noise of mean 0 and standard deviation σ was applied to the soma of each neuron at each time step of duration $dt = 1$ ms throughout the entire session.

For Figure 6, the simulated network is based on the one used in Figure 4F. In Figure 6B, $\sigma = 3$, 6, or 9, and the external stimulus amplitude I_A = 10, 15, or 20. 10 trials were conducted for each condition. For each trial, the memory amplitude was estimated by averaging the firing rates of the 10 highest firing neurons for a memory period of 10,000 ms.

In Figure 7A,B, we simulated diffusion of the bump location under noise for different bistable ranges, starting from the zero-width, step function case with $T_u = T_d$ and increasing the bistable range $T_u - T_d$. Increasing T_u changes not only the bistable range, but also the strength of synaptic input to each dendrite that is required during the encoding period to achieve a given amplitude of memory activity. Therefore, to approximately keep the memory activity the same for simulations with different bistable ranges, we changed the simulation protocol so as not to start each simulation from zero activity. Instead, we first initialized the network by simulating a network with $T_u = T_d = 3$, $I_A = 15$, and no noise for encoding and memory periods of duration 1000 ms. The network activity state at the end of this simulation then served as the initial condition for the shown simulations. From this initial condition, we then tested the effect of applying noise with standard deviation $\sigma = 10$ for a period of 10,000 ms for different values of T_u and T_d. In Figure 7A, the blue trace corresponds to $T_u = T_d = 3$, the orange trace has $T_u = 3.8$ and $T_d = 2.2$, and the green trace has $T_u = 4.6$ and $T_d = 1.4$. For each trace, we recorded the variances of the memory locations, estimated by the center of mass of neural activity, and averaged over 400 trials. In Figure 7B, the final location variance is plotted

against the bistable range, with $T_u = 3 + 0.2s$ and $T_d = 3 - 0.2s$ for integers s from 0 to 10. In Figure 7C, we perturbed the connectivity to add a small asymmetric component. This was done by increasing the weights of clockwise-directed connections by 5% and decreasing the weights of counter-clockwise-directed connections by 5%. All other parameters were identical to those used for the blue and green traces in Figure 7A,B.

For Figures 6 and 7, the shown plots begin at 300 ms after the offset of the stimulus to allow the transient remaining effects of the stimulus to decay away.

Implementations of numerical simulations

For all simulations, we integrated the differential equations using the forward Euler's method with a time step of 0.01 ms for the spiking model and 1 ms for the rate-based ring model, except for Figures 4B,F, 5, and 6A, where 0.01 ms was used to better approximate the simulated multistable band.

Code accessibility

Code was written in Python version 3.9.16 and is available in Github [44].

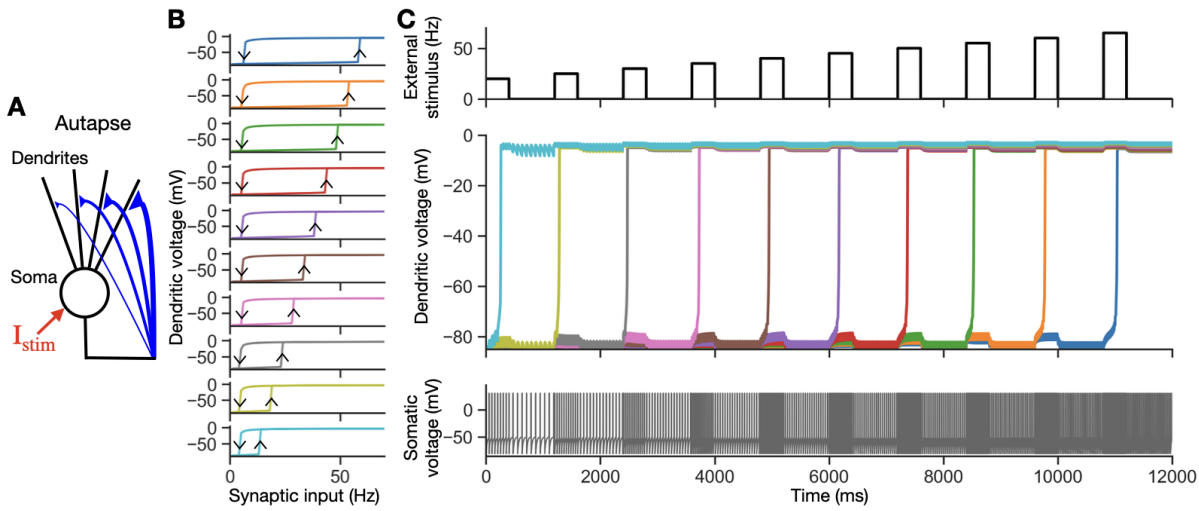

Figure 2: Maintenance of graded firing rates in the spiking autapse model. **A**, Model architecture. The model consists of an integrate-and-fire soma that projects to ten conductance-based dendritic compartments with bistability based on the voltage-dependent properties of the NMDA and Kir channels (see Materials and Methods). Each compartment receives synaptic input with a different weight (the maximal NMDA receptor conductance, $\bar{g}_{NMDA,i}$), so that different dendritic compartments are recruited sequentially with increased amplitude of synaptic input. The external stimulus I_{stim} (red) goes into the soma. **B**, Hysteretic input-output relations of the ten dendrites. To characterize the bistable input-output relations of each dendrite, the synaptic inputs from the soma were removed and replaced by periodically firing, externally applied synaptic inputs that stepped across a range of different frequencies (x-axis). The voltages shown represent averages over 1000 ms of recording. Arrows in the top panel indicate the directions of the hysteretic response. Top to bottom panels show the 10 dendrites, ordered by increasing NMDA conductance $\bar{g}_{NMDA,i}$ (or equivalently, decreasing up threshold). **C**, Memorization of different amplitudes. For the model in **A**, an external stimulus spike train to be encoded by the circuit is briefly applied, followed by a memory period with no stimulus. This is repeated 10 times for successively increasing external stimulus frequencies (top). Dendrites are activated during the encoding periods, starting from the one with the largest $\bar{g}_{NMDA,i}$, and self-sustain their activity during the memory period (middle). Somatic activity, which represents the memorized stimulus during the memory period, reflects the number of activated dendrites (bottom).

Results

Storage of the graded amplitude of a stimulus in an autapse circuit with bistable dendrites

SPIKING AUTAPSE CIRCUIT MODEL WITH MULTIPLE BISTABLE DENDRITES

We constructed a spiking autapse circuit model to demonstrate how a single biophysically based neuron with multiple conductance-based dendritic compartments can robustly sustain graded amplitude activity. The model neuron consists of an integrate-and-fire somatic compartment and 10 conductance-based dendritic compartments (Fig. 2A, Materials and Methods). Spiking output from the soma projects back onto each dendritic compartment through NMDA and AMPA receptor-mediated input, providing recurrent self-excitation. Due to the voltage-

dependent activation kinetics of the NMDA receptor and the inward rectifier potassium (Kir) channel, each dendritic compartment exhibits a bistable voltage response to synaptic input, similar to that described previously [43] (Fig. 2B). Briefly, in the absence of synaptic input, the hyperpolarization-activated Kir conductances maintain the dendritic voltage at a strongly hyperpolarized value. As the presynaptic firing rate of the neuron increases from zero, the fast excitatory postsynaptic potentials mediated by the AMPA receptors increasingly activate the slower, voltage-dependent NMDA conductances. Additionally, postsynaptic depolarization reduces the hyperpolarizing effect of the Kir conductance. When the input firing rate crosses an approximate threshold, there is a nonlinear increase in voltage up to a plateau level that is caused by the self-sustaining positive feedback loop between voltage activating the NMDA channels and the NMDA channels increasing the voltage (Fig. 2C, middle). This elevated voltage can then be sustained even for much lower levels of synaptic input. Different dendritic compartments have different synaptic input thresholds for flipping to this elevated, "up" voltage state, with the threshold level determined by the maximal conductance of the NMDA channels (Fig. 2B, dendrites ordered by level of their maximal conductance). The "down" threshold for flipping down the voltage is predominantly determined by the interspike interval at which NMDA-mediated synaptic inputs no longer summate sufficiently to maintain an elevated voltage, and thus is relatively insensitive to the level of NMDA conductance.

Due to the successive recruitment of dendritic compartments with different thresholds, the resulting circuit can sustain many different levels of persistent neural activity that can be used to store a graded representation of the amplitude of a stimulus. When a stimulus is presented (Fig. 2C, top; each pulse represents a separate stimulus), this leads to a transient increase in the firing rate of the neuron. This transient increase in firing then leads to the activation of those dendrites whose activation thresholds are crossed (Fig. 2B). When the stimulus turns off, the activated dendrites can maintain the firing rate of the neuron at a level determined by the number of dendritic compartments activated during the transient stimulus (Fig. 2C). In this manner, the circuit maintains a graded memory of the amplitude of the stimulus, with the number of different amplitude values that can be maintained equal to the number of dendritic compartments. Further, the maintenance of the memory activity is robust due to the bistability

of the dendritic response, for which small changes in firing rate cause minimal changes in dendritic voltage (Fig. 2B). In the following section, we use a simplified, firing rate based model to rigorously characterize the core computational mechanism by which this robust storage of graded stimulus values occurs.

FIRING RATE BASED AUTAPSE CIRCUIT MODEL WITH MULTIPLE BISTABLE DENDRITES

We constructed a firing rate based, multi-dendrite autapse circuit model that captures the essence of the spiking model while permitting rigorous, analytic understanding of the memory storing mechanism. The model consists of a neuron receiving input from two sources: an external stimulus I_{stim} whose amplitude is to be stored in the autapse circuit, and N_d bistable dendritic compartments that receive recurrent synaptic input driven by the firing of the neuron. The dynamics of the firing rate r obeys:

$$\tau \frac{dr}{dt} = -r + \sum_{i=1}^{N_d} D_i(r) + I_{stim}, \tag{15}$$

where the first term represents the intrinsic decay of the neuronal firing rate, with time constant τ, and the bistable response of each dendritic compartment is given by

$$D_i(r) = \begin{cases} \beta\Theta(w_i r - T_u), & \text{activation from down to up state,} \\ \beta\Theta(w_i r - T_d), & \text{deactivation from up to down state.} \end{cases} \tag{16}$$

Here, w_i gives the strength of the synaptic connection onto the ith dendrite, and β gives the contribution of each activated dendrite to the neuron's firing rate. T_u and T_d give the threshold values of input to the dendrite, $w_i r$, required to flip the dendrite to its up or down state, respectively. Equivalently, the dendritic input-output relation can be characterized by the threshold values of the firing rate required to activate or deactivate the dendrite, $r^{th}_{u,i}$ and $r^{th}_{d,i}$ (Fig. 3A, left). Summing together the individual bistable input-output relations of each of the dendrites gives a multistable, staircase-like relation for the total dendritic feedback to the neuron, with the corners of the stairsteps defined by the set of $r^{th}_{u,i}$ and $r^{th}_{d,i}$ values (Fig. 3A, right; Fig. 3B, gray boxes).

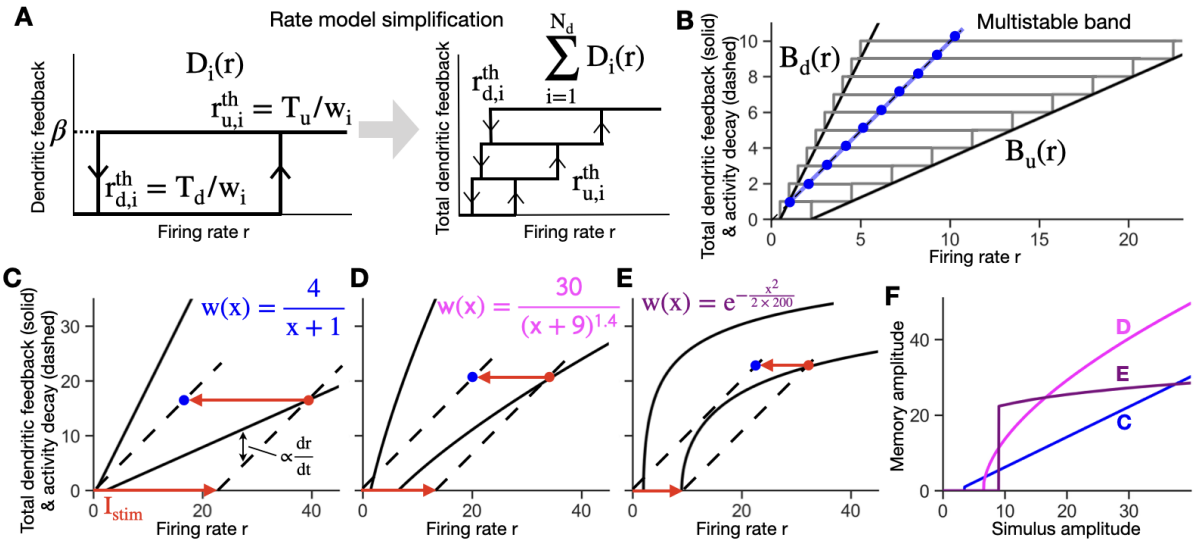

Figure 3: Rate-based autapse model for storage of graded amplitude stimuli. **A, B**, Rate based model captures the key behavior of the spiking model of Figure 2B. **A**, Left, bistable input-output relation $D_i(r)$ of a dendrite i. The input is the somatic firing rate, which feeds back to the dendrite through the self (autaptic) connection. $D_i(r)$ is hysteretic with jumps at up and down thresholds, whose values are inversely related to the synaptic weight w_i. Right, summed dendritic output to the soma from all dendrites (3 shown). Each rectangle represents the contribution from one dendrite. Dendrites with smaller w_i values have smaller thresholds and are plotted at the bottom. **B**, Multistable band in the discrete or continuous case. Multiple dendrites form a band with right and left edges B_u, B_d consisting of up and down thresholds, either in discrete steps (gray) or the continuum limit (black). The dashed line represents the decay of firing rate activity (magnitude of term –r in Eq. (15)). Intersections of the line and multistable band (the blue line in the continuous case or blue dots in the discrete case) represent points at which the total dendritic feedback offsets firing rate decay to lead to persistent firing (i.e., memory) states in the absence of an external stimulus. **C-E**, Different weight functions lead to different shapes of the multistable band and different responses to an external stimulus. **C**, Linear band derived from a power law decay weight function with power $p = 1$ (Mathematical Appendix). During the encoding period, an external stimulus I_{stim} shifts the dashed line (representing the combined effect of neuronal decay and the external stimulus) rightwards, with activity stabilizing at the red dot. The rate dr/dt of firing rate increases is proportional to the difference between the dashed line and rightward band edge B_u. In the memory period following the offset of the external stimulus I_{stim} (red leftward arrow), memory activity stabilizes at the blue dot. **D**, Same as **C**, but for a power law weight decay with $p = 1.4$. **E**, Same as **C**, but with an example Gaussian weight function. For this case, when the external stimulus barely exceeds the smallest up threshold, the memory activity abruptly jumps to a high level, amplifying the external stimulus. **F**, The form of the weight function determines the shape of the relationship between the stimulus amplitude and the amplitude of stored memory activity. Traces show the mappings between stimulus and memory amplitude for each weight function used in **C-E**. The blue trace represents linear encoding of stimuli in memory (Mathematical Appendix).

For the neuron to persistently maintain firing activity in the absence of a stimulus (i.e., have $dr/dt = 0$ when $I_{stim} = 0$), the recurrent feedback provided through the dendrites must offset the intrinsic leakiness of the neuron. Mathematically, from Eq. (15), this occurs when the total dendritic feedback equals the intrinsic neuronal decay term, $\sum_{i=1}^{N_d} D_i(r) = r$. Graphically, this can be depicted by plotting each of these terms as a function of the firing rate of the neuron

(Fig. 3B). The total dendritic feedback term $\sum_{i=1}^{N_d} D_i(r)$ is represented by a multistable band extending between left and right edges $B_d(r)$ and $B_u(r)$, respectively (Fig. 3B, rectangles). The intrinsic decay of the firing rate (corresponding to the term –r in Eq. (15)) is given by a line of unity slope (Fig. 3B, blue). The points of intersection of the multistable band and this line (Fig. 3B, blue dots) give the possible stable values of persistent firing activity that can be used to store different stimulus amplitudes. As can be seen, the occurrence of many possible stable values of firing activity does not require fine tuning – due to the width of the bistable band, weights could be changed over a wide range, warping the band, without losing the points of intersection [13, 28, 29]. This contrasts with classic graded amplitude circuits lacking bistable dendrites in which the feedback *band* collapses into a feedback *line* so that fine tuning is required for the circuit to be able to maintain many different persistent firing rates [1].

For the circuit to encode a stimulus into memory, dendrites activated by the external stimulus must remain activated during the subsequent memory period. Graphically, the ability of an external stimulus to activate dendrites can be depicted by shifting the unity-slope line that represents the decay term –r in Eq. (15) to the right by an amount I_{stim} (Fig. 3C, rightward pointing red arrow). Dendrites are activated up to the point where the multistable band first intersects the shifted unity-slope line, with the rate of increase of firing activity proportional to the difference between the edge of the multistable band and the shifted line (Fig. 3C, dr/dt). Subsequently, when the external stimulus is turned off, the system relaxes to a nonzero value (Fig. 3C, blue dot) that reflects the amplitude of the previous stimulus.

The shape of the multistable band is determined by the distribution of synaptic weights w_i in the circuit. For a neuron with many dendrites, this weight function can be approximated by a continuous function w(x), where the continuous value x replaces the discrete dendrite index i (Materials and Methods). Different weight functions give rise to different shapes of the multistable band (Fig. 3C-E). These different band shapes result in different relationships between the external stimulus amplitude and the amplitude of memory-storing activity (Fig. 3F). For weight functions with a power law decay of the form $w(x) = a/(x + b)$, as in Figure 3B,C, the band edges are linear, giving rise to a linear relationship between the amplitude of memory activity and the stimulus amplitude (Fig. 3F, blue; Mathematical Appendix). For

the weight function used in Figure 3D, the relationship between the stimulus amplitude and memory activity is continuous but with a nonlinear, concave shape (Fig. 3F, gray). As a result, the circuit devotes a relatively higher resolution in memory (larger fraction of its firing rate range) to smaller stimulus values, which could be useful if a task demands more sensitivity to lower amplitude stimuli. In Figure 3E, the stronger nonlinearity in the band shape leads to a discontinuous jump in the memory amplitude at a threshold level of the input stimulus, followed by a very robust, largely saturated response (Fig. 3F, purple). A feature of this connectivity structure is that it amplifies small stimulus values into large firing rates during the memory period. General mathematical formulas for the shape of the multistable band and the stimulus amplitude-memory amplitude relation, for a given weight function, are derived in the Materials and Methods and Mathematical Appendix.

Working memory in the rate-based ring model with bistable dendrites

Building on the insights gained from the autapse model, we constructed a rate-based ring model consisting of neurons with multiple bistable dendrites. As in classic ring models, neurons in the model are connected through local excitation and broader inhibition (Fig. 4A, left). Each of the N neurons in the network receives recurrent excitatory inputs through bistable dendrites (Fig. 4A, right) and recurrent inhibitory and feedforward stimulus inputs through direct connections onto the soma. The resulting equation for the firing rate r_i of the ith neuron in the model is:

$$\tau \frac{dr_i}{dt} = -r_i + \sum_{j=1}^{N} D_{ij}(r_j) - w_{inh} \sum_{j=1}^{N} r_j + I_{stim,i}, \tag{17}$$

where, as in the autapse model, the first and last terms on the right-hand side represent the intrinsic decay of the neuron and the external stimulus input into this neuron, respectively. The third term corresponds to recurrent inhibitory input not present in the autapse model. This was modeled as being all-to-all with strength w_{inh}, but the same qualitative behavior can be obtained with non-uniform lateral inhibition. The second term represents the recurrent excitatory inputs and was governed by dendritic input-output relations of the form:

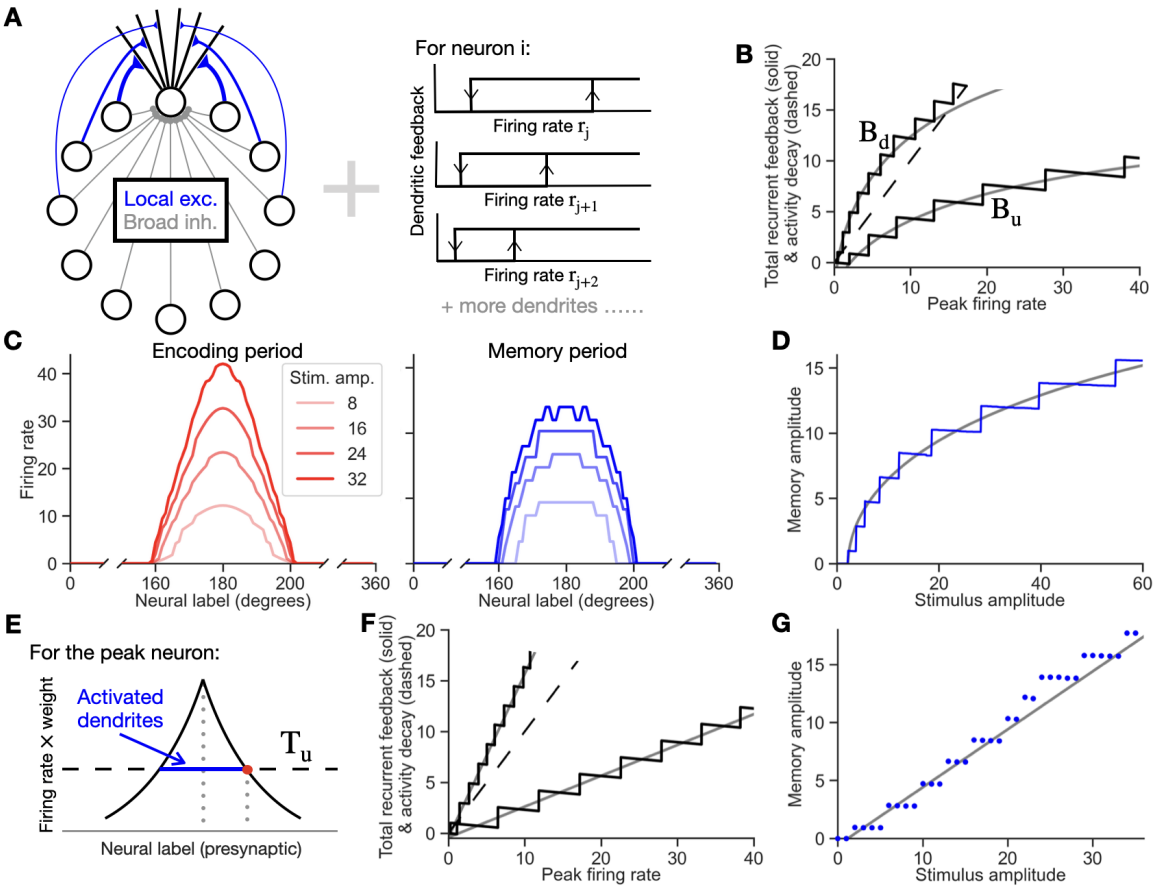

Figure 4: Simultaneous storage of stimulus location and amplitude in the ring model with dendritic bistability. **A**, Network architecture. Left, each neuron receives local excitatory input through connections onto separate dendrites and broad (uniform) inhibitory input through connections onto the soma. Right, each dendrite is bistable. Dendrites receiving inputs through larger synaptic weights (bottom) have lower thresholds for presynaptic neuron firing to flip the dendrite to its up state. **B**, Simulated multistable feedback band for peak firing neuron (black) and approximate, analytically calculated multistable feedback band for the whole ring model network (gray). See Materials and Methods for details. **C**, Storage of stimulus amplitude at a given location. Left, steady-state pattern of activity during the encoding period for a stimulus with the shown amplitude (legend) and location centered at 180 degrees. Right, stable pattern of persistent activity during the memory period. Each stimulus amplitude is faithfully encoded during the memory period. **D**, Mapping between stimulus and memory amplitude. Blue, simulations. Gray, approximate, analytically calculated relationship (see Mathematical Appendix for derivation). **E**, Schematic illustrating how different presynaptic inputs contribute to a postsynaptic neuron's level of activity during the encoding period (see Mathematical Appendix). The effective input a postsynaptic neuron receives at each dendrite is given by the product of the presynaptic neuron's firing rate and synaptic weight onto the dendrite. If this product exceeds the up threshold value T_u (dashed line), the dendrite is activated. The total number of activated dendrites of a neuron equals the length of the blue segment. The example shown is for the peak neuron at the center of a local stimulus pattern. **F**, Multistable band for a network with weights chosen to produce a linear relation between stimulus and memory amplitude. Black, simulated multistable band for the peak neuron. Gray, approximate analytically calculated band for the whole network. **G**, Mapping between stimulus and memory amplitude for weights analytically calculated to approximate a linear relation, as in **F**.

$$D_{ij}(r_j) = \begin{cases} \beta\Theta(w_{ij}r_j - T_u), & \text{activation from down to up state,} \\ \beta\Theta(w_{ij}r_j - T_d), & \text{deactivation from up to down state.} \end{cases} \tag{18}$$

Here, w_{ij} gives the strength of the input arriving onto neuron i from neuron j, which we assume projects to the jth dendrite of neuron i. This assumption that each presynaptic excitatory input arrives on a devoted dendrite was made primarily for analytic tractability, but could be realized biophysically if each of the dendritic compartments in the model corresponds to a dendritic spine. Although the model is written formally as if each neuron has N dendritic compartments, compartments with synaptic strengths w_{ij} too weak to activate the corresponding dendrite could be removed with no change in model output. For the model instantiations considered below, the resulting effective number of dendritic compartments was between 15 and 20.

Intuitively, this model can be seen as a spatially extended version of the autapse circuit model. The autapse circuit directly provides feedback onto its multiple bistable dendrites so as to be able to maintain stable firing at multiple levels. Analogously, in the ring model, a group of active neurons together provide feedback onto themselves to maintain a pattern of neuronal firing that can achieve multiple stable levels. Further, as in traditional ring models, the symmetry of the model's connectivity implies that it can stably maintain such a multistable bump of activity at any given location. This intuition, and the close relation of this network's behavior to that of the autapse circuit model, is formalized more quantitatively below and in the Mathematical Appendix.

For activity to be maintained in memory following the offset of a transiently presented stimulus, each neuron must achieve a stable, nonzero firing rate ($dr_i/dt = 0$) in the absence of external stimulus input ($I_{stim,i} = 0$). Similar to the autapse model, this requires that the total recurrent feedback onto each neuron (second and third terms of Eq. (17)) balances the intrinsic decay of the neuron's firing rate (first term of Eq. (17)). This can be depicted graphically for each neuron by plotting the total recurrent feedback as a function of the neuron's firing rate (Materials and Methods). As in the autapse model, the feedback forms a multistable band, with left and right band edges defined as B_d and B_u, respectively (Fig. 4B). The negative slope of each step of the total recurrent feedback band is due to the recurrent inhibitory input that was not present in the autapse model. For a given neuronal firing rate and level of steady state recurrent

feedback, persistent activity can be maintained by this neuron at the point at which the recurrent feedback band intersects the unity-slope line corresponding to the intrinsic decay of firing (Fig. 4B, dashed line, corresponding to the term –r in Eq. (17)).

A network simulated with this architecture can successfully maintain memory activity that reflects the amplitude of an externally presented stimulus. During the encoding period, when the stimulus is present, neural activity reflects both the direct influence of the external stimulus and the recurrent feedback mediated by the bistable dendrites and recurrent inhibition (Fig. 4C, left, shown for a stimulus that entered the network through a Gaussian shaped input profile). Following the offset of the stimulus, the recurrent interactions between the neurons maintain activity at a firing rate that reflects the amplitude of the external stimulus (Fig. 4C, right). The relationship between the amplitude of the stimulus and the amplitude of the memory activity encoding this stimulus is shown in Figure 4D (blue line).

The graphical construction described above (Fig. 4B) only characterizes the dynamics of a single neuron at a time, rather than the collective behavior of the network as a whole. Because the shape of the band of a given neuron depends on the activity of the other neurons in the network, this single-neuron band is generally insufficient to describe the whole-network behavior in a manner comparable to the analysis of the autapse circuit. However, with some simplifying assumptions, we found that we could define a single multistable band that well-approximated the behavior of the entire network and enabled us to predict the memory amplitude vs. stimulus amplitude relation (such as that in Figure 4D) for a given set of network parameters. As described more fully in the Mathematical Appendix, if we assume that the responses of the network during the encoding period and the memory period are approximately the same shape as that of the external stimulus profile, then the equations for the amplitude of the bump of activity reduce down to equations mathematically analogous to those for the amplitude of activity in the autapse model, modified by a factor accounting for the presence of recurrent inhibition. The resulting analytically calculated band for the whole network has a shape (Fig. 4B, gray lines) that closely approximates the simulated multistable band for the neuron at the peak of the bump of activity (Fig. 4B, black lines), up to deviations that occur at the very highest sustainable firing rates (around approximately 15 Hz in Figure 4B), due to the activity bump notably changing

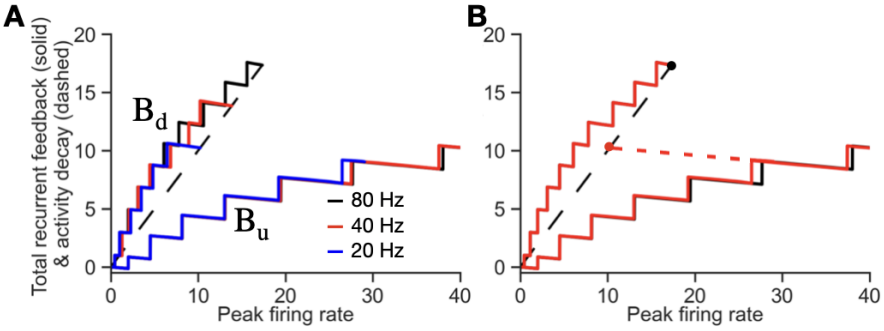

Figure 5: Similar multistable bands under different stimulus amplitudes and presentation patterns. To demonstrate the fidelity of the approximate analytic characterization of the ring model dynamics as being determined by a single multistable band independent of the amplitude of the external stimulus, here we show how little the simulated bands vary under a broad range of different external stimulus conditions. **A**, Varying the stimulus amplitude over a broad set of values. **B**, Applying two successive stimuli, with no offset between them. First, a smaller stimulus amplitude of 20 is applied, resulting in a smaller memory amplitude (red dot). Then, a larger stimulus amplitude of 80 is applied, resulting in a larger memory amplitude (black dot). The black multistable band (hidden on the left side of the band by the overlying red band) is for a stimulus amplitude of 80 with no preceding stimulus, i.e. when starting from zero firing activity.

shape for very large stimuli. Away from this saturating limit, the analytically calculated memory amplitude vs. stimulus amplitude relation for the network (Fig. 4D, gray line) closely approximates that of the network simulation (Fig. 4D, blue line). Thus, the same fundamental mechanisms that precisely describe the amplitude-storing dynamics of the autapse circuit can also, to a very good approximation, be used to describe the amplitude-encoding dynamics of the ring model.

The analytic calculations enabled us to determine how the shape of the network's multistable band and the shape of the memory amplitude vs. stimulus amplitude relationship depend on the synaptic connectivity profile w(x) and the shape of the external stimulus profile R(x), where x is the distance between neurons along the ring (Mathematical Appendix). For example, Figure 4F compares the simulated (black) and analytically calculated (gray) bands for a weight profile derived to produce a linear-shaped band, and Figure 4G shows the resulting memory amplitude vs. stimulus amplitude relationship (gray: analytically calculated; blue: simulations). We note that the analytic model predicts that there is little dependence of the multistable band on the amplitude of the external stimulus (with small deviations resulting from the shape of the bump notably changing for large external stimuli). The simulations verified this prediction, showing that the shape of the band for a given network is quite consistent across a wide range of stimulus

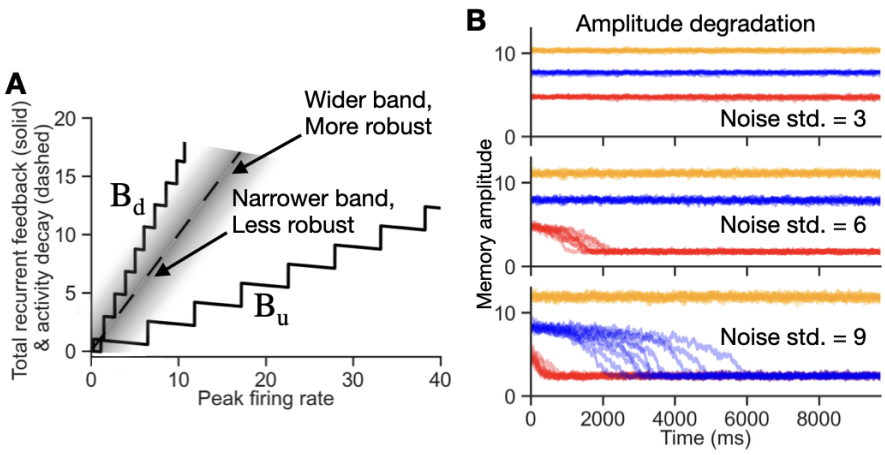

Figure 6: Robustness of memory amplitude to noise and amplitude drift. **A**, Schematic of the effect of noise on memory amplitude. Noise acts as a time-varying external stimulus to each neuron, whose effect can approximately be envisioned as fluctuating the dashed line, along which stable points lie in the no-noise case, horizontally left or right (gray cloud). A memory amplitude is stable if noise does not drive stable memory states across an edge of the band, i.e. if it does not flip dendrites on or off. The robustness to noise depends on the noise level and geometry of the multistable band, with wider portions of the band more robust to noise than narrower portions. **B**, Examples of amplitude drift. Demonstration of the concept sketched in **A** for simulations of the network of Figure 4F with 3 different noise amplitudes (top to bottom panels) for 3 different stimulus strengths (orange, blue, red traces: stronger to weaker stimuli, respectively). Ten simulated trials are shown for each condition, with memory amplitude measured by the average firing rate of the 10 most active neurons in the bump of activity. For small noise (top), all three memories remain stable for each shown stimulus amplitude. For moderate noise (middle), the memory responses with a lower stimulus amplitude (red) go to baseline activity. For larger noise (bottom), memory of the lowest amplitude (red) degrades quickly and that of the middle amplitude (blue) degrades more slowly. The nonzero baseline memory amplitudes following loss of the memory reflect positive fluctuations in firing rate driven by the noise, rather than any remaining storage of memory amplitude.

amplitudes (Fig. 5A). Further, if a second stimulus is applied while a first stimulus is still being stored, the resulting band for the second stimulus is minimally affected by the first stimulus (Fig. 5B, black is for a single large stimulus and red is for a small stimulus as in Figure 5A, followed by the large stimulus).

Robustness of the rate-based ring model with bistable dendrites

Traditional continuous attractor models for the storage of the amplitude [1,2] or location [3, 4] of a stimulus both require fine tuning of their connectivity to maintain analog representations and exhibit a diffusion of activity in the presence of noise that degrades the fidelity of the memory over time [45,46,47,48]. By contrast, model networks with bistable components can have greater robustness to perturbations in their storage of amplitude [13,28,29] or location [30] information. Here, we examine the robustness of the ring model with multiple bistable dendrites

to the simultaneous storage of both memory amplitude and location.

We first considered maintenance of the amplitude of a memory state in the presence of noise (Fig. 6). Mathematically, noise appears as a fluctuating external stimulus term in Eq. (17) for each neuron of the network. Thus, similar to the stimulus-driven external input (cf. Fig. 3C), the effect of such noise can be thought of as horizontally jiggling back and forth the dashed line along which stable memory states lie (Fig. 6A). Graphically, this suggests that the width of the multistable band determines the level of robustness to noise: if the standard deviation of the noise input (Fig. 6A, gray shadowing around the dashed line) is much less than the distance to either edge of the band, the memory amplitude is expected to remain stable, whereas if the standard deviation exceeds the distance to either edge of the band, then dendrites will flip on or off, and the network activity is expected to change. Specifically, if the left edge of the band is closer to the no-noise memory state locations (along the dashed line) than the right edge of the band, as in Figure 6A, then noise will more often turn off dendrites than turning them on, and the overall activity will be expected to decline over time. To test this idea, we injected a random white noise input to each neuron in the network and then monitored how well the network could maintain three different external stimulus amplitudes (Fig. 6A, colored lines). For moderate levels of noise, the resulting amplitudes of persistent neural activity are held nearly perfectly constant in memory, exhibiting no diffusion (Fig. 6B, top). When the noise is increased, the lowest amplitude memories degrade (Fig. 6B, middle and bottom) whereas the largest amplitude stimulus could still be robustly stored in memory.

We next considered the maintenance of the location of a stimulus in memory in the presence of noise or asymmetry in connectivity (Fig. 7). For a network without bistable dendrites (Fig. 7A, blue), the variance of the location of the bump of activity in the network in the presence of neuronal noise increased rapidly and approximately linearly over time, consistent with the well-documented diffusion of the memory location over time in classic bump attractor networks [45, 46]. By contrast, for networks with bistable dendrites, the diffusive drift was decreased or nearly completely eliminated (Fig. 7A, orange and green traces), with the amount of diffusive drift decreasing monotonically with the width of the bistable dendritic input-output relation (Fig. 7B). Further, networks with bistable dendrites (Fig. 7C, green trace) could quite stably

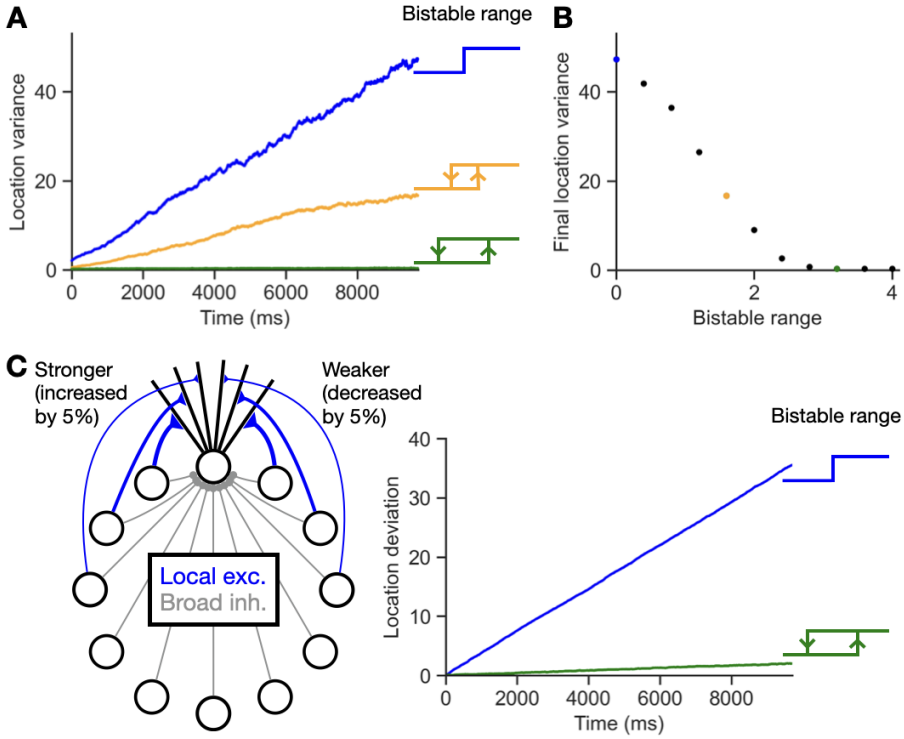

Figure 7: Robustness of memory location to noise with dendritic bistability. **A**, Dendritic bistability reduces the diffusion of the location of the bump of memory activity. Plots show the variance of the memory location over time for 400 trials. If each dendrite has a step function output, $T_u = T_d$, the diffusion is severe with a large location variance over time (blue). If the dendritic bistable range, $T_u - T_d$, is larger, the diffusion is reduced (orange or green). For all figure panels, the parameters defining the band of bistability were identical to those shown in Figure 6A except for the values of T_u and T_d (see Materials and Methods). **B**, Reduction of location diffusion. The variance of the location at the last time step of **A** is recorded for the shown bistable ranges. As the bistable range increases, the location diffusion decreases and eventually becomes negligible when the bistable range is wide enough that noise very rarely affects the dendritic states. Colored points correspond to colored traces in **A**. **C**, Robustness to asymmetry of ring model recurrent connectivity. Left, the ring model connectivity was perturbed by increasing clockwise directed connections by 5% and decreasing counter-clockwise directed connections by 5%. Right, breaking of the symmetry of the connectivity caused strong drift in the bump location in a model with no dendritic bistability (blue) but very little drift in the model with bistable dendrites (green).

maintain a bump of activity at a given location in the presence of small asymmetries in network connectivity, whereas the same amount of asymmetry led to rapid drift in the bump location for networks without bistability in their dendrites (Fig. 7C, blue trace). These results reflect that the width of each dendrite's bistable relation determines its tolerance to noisy input or to perturbations of weights that alter input currents, both for storing the amplitude (Fig. 6) and location (Fig. 7) of a stimulus. Thus, neurons with bistable dendritic compartments can provide a network with a highly noise-tolerant substrate for working memory storage.

Discussion

Biological systems are capable of remembering both the amplitude and location of stimuli. Previous working memory models have either been capable of storing only amplitude or only location information, or have been non-robust to noise and perturbations of synaptic weights. Here we have demonstrated how amplitude and location may simultaneously and robustly be encoded through the inclusion of bistable dendritic compartments in neurons. For a sufficiently large number of bistable dendrites in each neuron, the model can robustly maintain an approximately continuous amplitude at any given location. We show through simulations and analytic calculations how this robust storage of amplitude information can be understood geometrically as resulting from a multistable band of activity within which activity can be robustly maintained without fine tuning, and how the shape of this band is determined by the form of the synaptic weight distribution.

Biophysical implementation of the model

The fundamental building block that provides robustness to our model is dendritic bistability. Following previous work [41, 42, 43], we show how such voltage bistability can be realized through the voltage-dependent activation properties of NMDA receptors and Kir conductances. The key feature of these ionic conductances is their voltage-dependent activation, which mediates positive feedback in the cellular voltage dynamics: the depolarization-dependent activation of an inward conductance such as the NMDA channel robustly maintains the elevated voltage state, while the hyperpolarization-dependent activation of an outward conductance such as the Kir channel robustly maintains the low-voltage state. Other ion channels such as voltage-sensitive Ca^{2+} conductances also have this property and have been demonstrated to contribute to cellular or dendritic plateau potentials [32, 40, 49].

NMDA receptors particularly have been demonstrated to be critical to working memory function [50, 51, 52] and have been a major component of working memory models [12]. However, it is still unclear whether this is simply due to their slow kinetics, which makes memory activity more stable [53], or whether it is also related to their ability to induce dendritic bistability. Physiological recordings have shown that NMDA receptors in dendrites can cause pro-

longed voltage plateaus lasting hundreds of milliseconds [49, 54, 55], but more direct evidence is needed to determine whether dendritic plateau potentials can be sustained for seconds in the presence of persistent synaptic input during working memory tasks. Even if dendritic compartments only sustained prolonged plateau potentials, rather than maintaining truly bistable states, such plateau potentials could nevertheless provide a long intrinsic timescale for working memory that is then extended by network-level feedback interactions [2, 13].

A critical question is how many bistable states can be independently maintained in a single neuron. Classic modeling work based upon morphologically reconstructed neurons [56, 57, 58, 59] suggests that neurons are capable of maintaining several tens of approximately independent voltage compartments, consistent with experiments suggesting highly compartmentalized function [49, 55, 60, 61, 62, 63]. Very many compartments could be maintained independently if the functional voltage subunit was a dendritic spine, as suggested by recent experiments [62, 63], with localization of bistable voltage compartments enhanced by the conditionality of NMDA receptor activation on presynaptic input activity [41].

Alternatively, the multistable single-cell dynamics studied here could be realized in a larger network in which either individual neurons [34, 35, 36, 37, 38, 39, 40] or clusters of neurons [13] provided the fundamental bistable subunit. In this case, the successive recruitment of bistable dendrites in our model would need to be replaced by the successive recruitment of bistable neurons connected in an appropriate manner to functionally produce the stable, graded-amplitude activity demonstrated here. For example, a literal mapping of our model to a network architecture could be generated if, instead of dendrites projecting to a soma, neurons projected to a common neighbor that acted like a hub, with this hub neuron then providing output to other, similarly arranged neuronal groups arranged in a ring-like architecture. Such an architecture would presumably need to be learned through training. Towards this direction of learning connections between bistable units, recent work inspired by gated recurrent units (GRU) from machine learning have suggested how the synaptic connectivity of large networks of bistable units may be trained to solve tasks requiring memory to be stored in neural activity over long-timescales [64].

Robustness to noise and sensitivity to external stimuli

Our model shows a trade-off between the sensitivity to small stimuli and the robustness to noise. A wide bistable range in dendrites makes memory robust to noise in neural activities. However, this comes at the cost of reduced sensitivity to the encoding of small external stimuli [28], as a small stimulus may not be able to flip on dendrites and thereby encode information into memory.

This trade-off between robustness to noise and sensitivity to small stimuli might not be inevitable. At least in experiments studying the maintenance of head direction coding in Drosophila [65, 66], small head velocity stimuli can change the head direction value that is stored in memory. One possibility is that earlier stages of processing transform small stimuli to sufficiently large amplitudes that they can be encoded in the network [67]. Similarly, if spikes from external stimuli arrive in a correlated or synchronous manner, then each correlated spiking event corresponding to a stimulus may be sufficiently large to be encoded in memory. This would require that noise, by contrast, arrives in an uncorrelated manner, so that it does not disrupt encoding by bistable dendritic units. Alternatively, the noise robustness/small stimulus sensitivity tradeoff might be avoided if the dendritic compartments can flexibly switch between bistable and non-bistable input-output relations. This might be accomplished if, for example, the external stimuli to a network not only provide excitatory inputs to be encoded by the network but also simultaneously mediate disinhibition of dendritic compartments, changing their current-voltage relationships to a non-bistable regime. Such disinhibition enabling the occurrence of local regenerative dendritic events has been observed to occur in the hippocampal CA1 region, with lateral entorhinal cortical inputs providing disinhibition of CA1 pyramidal neuron dendrites [68, 69]. Altogether, these mechanisms suggest how networks may be able to take advantage of the robustness to noise and perturbations provided by bistable processes while also remaining sensitive to small stimuli.

Encoding of stimulus amplitude, width, or uncertainty in the multistable ring network

Previous ring models have primarily focused on the encoding of location information while neglecting the ability of networks to encode graded information at a given location. Such graded

amplitude encoding is the essence of line attractors for parametric working memory. However, line attractors generically require fine tuning to generate the continuum of fixed points along the attractor basin [2, 3, 5, 13]. A small set of previous studies have produced ring networks that successfully maintain amplitude information [23, 24, 25, 26], but like classic line attractors, these networks all involved a fine tuning of synaptic connectivity. Alternatively, a recent study showed that a discrete set of amplitudes could be encoded in a ring network with oscillatory activity [27], but this network had a tradeoff between the degree of required tuning and the discreteness of the activity states, such that storing a larger set of amplitude levels required a larger degree of fine tuning. Building on previous work constructing robust, not fine-tuned line attractor models through the use of multiple bistable units [13, 28, 29], we have shown how the same principles enable the construction of a network that robustly stores both amplitude and location information. In addition, as in previous ring models with cellular bistability [31], our model can store stimuli of different widths, because the width of the sustained bump of activity in our models approximately corresponds to the width of the external stimulus (Fig. 4C and Mathematical Appendix).

The amplitude of a pattern of neural activity can be interpreted as encoding information in two different manners: as the intensity of a stimulus at a given location or as the precision of the location. Here, we have largely framed our discussion of the ring model in the former manner, as encoding the location and intensity of a stimulus. Alternatively, Bayesian models of sensory coding [26,70,71,72,73] show how the amplitude of neural firing can represent the precision of a location, with higher firing rates indicating higher precision. Such a representation of precision enables Bayesian computations to be performed by summations of firing rates [19,20]. Building on this Bayesian framework, recent work has shown how a ring model with slowly decaying amplitude of activity can encode uncertainty about location or orientation [26]. This encoding of uncertainty enabled the network to perform accurate Bayesian inference in tasks such as path integration and evidence accumulation, with suggestions for how this form of coding may apply to the head direction system. Within the Bayesian framework, our dendritic multistability model provides a mechanism by which uncertainty about location or orientation could be robustly maintained in working memory. Thus, dendritic bistability could provide an intrinsic cellular

mechanism that enhances the robustness of computations underlying both working memory and Bayesian inference.

Mathematical Appendix

Here we provide the analytical derivations that underlie the exact or approximate characterizations of the dynamics and steady-state responses of the rate-based model networks. First, we present the calculations that exactly relate the memory-storing properties of the autapse circuit to its weight function. Next, for the ring model, we use a simplifying assumption to derive its approximate multistable band. We then show how, under this simplifying assumption, the mathematical form of this band and the relation between the memory amplitude, stimulus strength, and weight function reduce to a form similar to that of the simpler autapse model.

Determination of the relationship between the memory amplitude, the stimulus strength, and the weight function in the rate-based autapse model

Here, we derive relationships between the steady-state amplitude of memory activity r_m, the stimulus strength I_{stim}, and the weight function w(x). Specifically, below we derive: (1) the form of the weight function w(x) required to achieve a given stimulus-memory amplitude relationship,

$$r_m = f(I_{stim}), \tag{19}$$

and (2) for a given weight function w(x), the function f().

To proceed, we make two assumptions. First, we assume that the neuronal activity during the encoding period is monotonically non-decreasing, so that dendrites can only remain at their previous value or become activated, and that the activity reaches steady-state by the end of the encoding period. Thus, following the derivation of Eq. (12), the total number of dendrites, z_{tot}, that are active at the end of the encoding period is given by

$$z_{tot}(r_e) = \frac{B_u(r_e)}{\beta} = w^{-1}(\frac{T_u}{r_e}), \tag{20}$$

where r_e represents the firing rate at the end of the encoding period. Second, we assume that the down threshold T_d is low enough that no dendrites flip from their up to down state following the offset of the external stimulus, i.e. that the B_d edge of the band is always to the left of the unity-slope line along which fixed points of the dynamics lie (Fig. 3B-E). Thus, the steady-state activity during the memory period differs from that of the encoding period, r_e, only by removal

of the external stimulus I_{stim}:

$$r_m = r_e - I_{stim}. \tag{21}$$

Substituting Eq. (20) into Eq. (21) then gives a relationship for the weight function in terms of the memory period activity r_m:

$$w(z_{tot}) = \frac{T_u}{r_m + I_{stim}}. \tag{22}$$

The steady-state memory period activity r_m is given by β times the number of activated dendrites,

$$r_m = \beta z_{tot}(r_e). \tag{23}$$

Substituting the above equation and the desired relationship Eq. (19) into Eq. (22) then gives:

$$w(z_{tot}) = \frac{T_u}{\beta z_{tot} + f^{-1}(\beta z_{tot})}. \tag{24}$$

Since this holds for any number $x = z_{tot}$ of dendrites that can be activated by an external stimulus of amplitude I_{stim}, we finally have an expression for the weight function required to produce a given relationship f() between the stimulus amplitude and memory amplitude:

$$w(x) = \frac{T_u}{\beta x + f^{-1}(\beta x)}. \tag{25}$$

The relations above can be rearranged to determine the relationship f() between the stimulus amplitude and memory amplitude for a given weight function w(x). From Eqs. (22) and (23),

$$w(\frac{r_m}{\beta}) = \frac{T_u}{r_m + I_{stim}}. \tag{26}$$

Thus, for a given weight function w(x),

$$I_{stim}(r_m) = \frac{T_u}{w(\frac{r_m}{\beta})} - r_m = f^{-1}(r_m), \tag{27}$$

and f() is given by the inverse of this relationship.

WEIGHT FUNCTION FOR A LINEAR MEMORY-STIMULUS RELATIONSHIP AND LINEAR MULTISTABLE BAND

For a linear stimulus amplitude-memory amplitude relationship $r_m = c(I_{stim} - I_{th})$, we have that:

$$I_{stim} = f^{-1}(r_m) = \frac{r_m}{c} + I_{th}, \tag{28}$$

which from Eq. (25) gives:

$$w(x) = \frac{a}{x+b}, \tag{29}$$

where $a = \frac{cT_u}{\beta(c+1)}$ and $b = \frac{cI_{th}}{\beta(c+1)}$. This relationship is shown in Figure 3F for $a = 4$ and $b = 1$. For this case, as shown in Figure 3C, the edges of the multistable band are also linear, as may be confirmed by substituting the above form of $w(x)$ into Eq. (20). For example, performing the calculation for the right side of the band gives:

$$B_u(r) = \frac{a\beta}{T_u} r - b\beta, \text{for } \frac{T_u b}{a} \leq r \leq \frac{T_u(N_d + b)}{a}. \tag{30}$$

Derivation of the approximate dynamics of the rate-based model with ring architecture

In this section, we provide analytic calculations of the approximate dynamics of the ring model. Unlike the autapse circuit, it is difficult to provide a simple, exact analytic characterization of the system dynamics described by Eq. (13). This is because, although each individual neuron's behavior for a given location of the memory-storing bump can be qualitatively conceptualized by a multistable band defined by the sum of its dendritic and recurrent inhibitory inputs, each dendrite in the ring model receives input of a different firing rate r_j. Therefore, the shape of such a multistable band does not just depend on the local excitatory weights w_{ij}, but also on the specific pattern of firing rates r_j. Thus, because r_j depends on the amplitude and shape of $I_{stim,i}$, the exact shape of the multistable band is not uniquely defined. However, as shown in the following sections, we can analytically derive an approximate multistable band (Fig. 4B,F) and relationship between stimulus amplitude and memory amplitude (Fig. 4D,G) that closely parallel the exact results from the autapse circuit and that well-approximate the corresponding simulation results.

CALCULATION OF HOW THE SHAPE OF THE WEIGHT FUNCTION DETERMINES THE MULTISTABLE BAND AND THE MAPPING BETWEEN THE EXTERNAL STIMULUS AND MEMORY AMPLITUDE IN THE RING NETWORK

Below, we first derive an approximate shape of the multistable band. We then determine approximate relationships between the amplitude of the external stimulus, the memory amplitude (i.e. steady-state firing rate of the neuron at the peak of the activity bump during the memory period), and the weight function. In particular, we show below how the derived relationships

tightly parallel and generalize those found for the autapse case, enabling the intuitions derived exactly for the autapse case to be related to the full ring model case.

During the encoding period, we assume for simplicity that neurons start from zero firing rate and only maintain or increase their firing rates when the external stimulus is presented. In this case, the term $D_{ij}(r_j)$ can be replaced in Eq. (13) by a step function, of height β, corresponding to the condition for activating a dendrite from its down to its up state:

$$\tau \frac{dr_i}{dt} = -r_i + \beta \sum_{j=1}^{N} \Theta[w_{ij} r_j - T_u] - w_{inh} \sum_{j=1}^{N} r_j + I_{stim,i}. \tag{31}$$

In the following, we work in the continuum limit, with the continuum neuron index x replacing the discrete index i. We denote the steady state firing rate that is approached during the encoding period by $r_e(x)$, which is given from the above equation by

$$r_e(x) = \beta \int_0^N \Theta[w(x-z) r_e(z) - T_u] dz - w_{inh} \int_0^N r_e(z) dz + I_{stim}(x), \tag{32}$$

where the excitatory weights $w(x-z)$ are a function of the shortest distance between x and z along the ring. A dendrite is activated if $w(x-z)r(z) > T_u$. Thus, the first integral gives the total number of activated dendrites. In Figure 4E, we illustrate schematically, for the maximally firing neuron of a bump of activity, how this integral can be graphically solved in response to a localized symmetric stimulus $I_{stim}(x)$. The total dendritic input ($w(x-z)r(z)$, black) to postsynaptic neuron x is plotted as a function of the presynaptic neuron index z. Activated dendrites are those (labeled in blue) that exceed the threshold value T_u (dashed line).

To derive an approximate multistable band and approximate relationships between the weight profile $w(x)$, external stimulus amplitude, and steady-state memory amplitude, we make two further assumptions. First, we assume T_d is low enough that no dendrites flip from their up to down state following the offset of the external stimulus (i.e. during the memory period). Therefore, the steady-state activity during the memory period $r_m(x)$ is:

$$r_m(x) = \beta \int_0^N \Theta[w(x-z) r_e(z) - T_u] dz - w_{inh} \int_0^N r_m(z) dz. \tag{33}$$

Note that the first integral in Eq. (33) is identical to the first integral in Eq. (32), indicating that the total number of activated dendrites is unchanged. Second, we assume that the stimulus $I_{stim}(x)$ and the steady-state firing rates during the encoding period $r_e(x)$ and during the memory

period $r_m(x)$ approximately have the same symmetric, unimodal (bump-like) shape $R(x)$, so that we can write $I_{stim}(x) = I_A R(x)$, $r_e(x) = E_A R(x)$, and $r_m(x) = M_A R(x)$. We define $R(x)$ to have a peak value of 1, so that I_A, E_A, and M_A give the amplitudes of the corresponding stimulus or activity profiles.

DERIVING THE APPROXIMATE SHAPE OF THE MULTISTABLE BAND

Given these assumptions, we first show that the network activity is governed by a multistable band of recurrent network feedback analogous to that derived for the autapse case (Eq. (12)). We focus our analysis on the neuron at the peak of the bump of activity, which we label without loss of generality as $x = 0$. From the paragraph above, the firing rates of all other neurons in the network can then be derived by scaling the activity of this neuron by the bump-shape factor $R(x)$. During the encoding period, the total dendritic feedback providing input to neuron $x = 0$, which we denote as $D_{tot}(E_A)$, is given from Eq. (32) by

$$\begin{aligned} D_{tot}(E_A) &= \beta \int_{-\frac{N}{2}}^{\frac{N}{2}} \Theta[w(z)R(z)E_A - T_u]dz \\ &= 2\beta \int_{0}^{\frac{N}{2}} \Theta[w(z)R(z)E_A - T_u]dz, \end{aligned} \tag{34}$$

where in the above we have used the symmetry of the weight matrix and the neural activity profile. Performing the integral analogously to the calculation for the autapse case then gives:

$$D_{tot}(E_A) = \begin{cases} 0, & \text{for } E_A < \frac{T_u}{w(0)}, \\ 2\beta W^{-1}(\frac{T_u}{E_A}), & \text{for } \frac{T_u}{w(0)} \leq E_A \leq \frac{T_u}{w(\frac{N}{2})R(\frac{N}{2})}, \\ \beta N, & \text{for } E_A > \frac{T_u}{w(\frac{N}{2})R(\frac{N}{2})}, \end{cases} \tag{35}$$

where $W^{-1}(x)$ denotes the inverse of the function $W(x) = w(x)R(x)$. Therefore, the total recurrent feedback at steady-state during the encoding period, which specifies the right side of the multistable band as a function of the amplitude of the bump of neural activity, is given from Eq. (32) by

$$B_u(E_A) = \begin{cases} -w_{inh}\kappa E_A, & \text{for } E_A < \frac{T_u}{w(0)}, \\ 2\beta W^{-1}(\frac{T_u}{E_A}) - w_{inh}\kappa E_A, & \text{for } \frac{T_u}{w(0)} \leq E_A \leq \frac{T_u}{w(\frac{N}{2})R(\frac{N}{2})}, \\ \beta N - w_{inh}\kappa E_A, & \text{for } E_A > \frac{T_u}{w(\frac{N}{2})R(\frac{N}{2})}, \end{cases} \tag{36}$$

where $\kappa = \int_0^N R(x)dx$ is a constant for a given shape $R(z)$ and depends only on the total integrated area of $R(z)$. Analogous reasoning can be applied to determine the left side of the

multistable band, B_d. In the case of no inhibition, $w_{inh} = 0$, this result is analogous to that found for the autapse model (Eq. (12)) with W acting analogously to the autapse weight function w and the factor of two being due to the range of x extending symmetrically from negative to positive values for the ring model. This analytical result for the band shape is compared to simulation results in Figure 4B,F.

DERIVING THE RELATIONSHIP BETWEEN THE MEMORY AMPLITUDE, THE STIMULUS STRENGTH, AND THE WEIGHT FUNCTION

We next derive the symmetric weight profile w(x) required to generate a given relationship,

$$M_A = f(I_A), \tag{37}$$

between the external stimulus amplitude and memory amplitude. Our derivation closely parallels that shown above for the autapse model.

First, recall that the memory amplitude M_A is defined as the firing rate of the neuron at the peak of the bump of activity during the memory period, and that this peak firing rate is determined by the total number of dendrites, z_{tot}, that are active at the end of the encoding period. For the nonzero and not saturated regime of Eq. (35),

$$z_{tot}(E_A) = \frac{D_{tot}(E_A)}{\beta} = 2W^{-1}(\frac{T_u}{E_A}). \tag{38}$$

z_{tot} is illustrated as the length of the blue segment in Figure 4E. Defining $z_{max} = z_{tot}/2$ as the index of the last activated dendrite (Fig. 4E, abscissa value of red point), from Eq. (38) we have that:

$$w(z_{max})R(z_{max})E_A = T_u. \tag{39}$$

To derive an expression for the memory amplitude M_A in terms of the stimulus strength I_A, first, we show how the steady-state amplitude of the bump at the end of the encoding period, E_A, is related to I_A. During the encoding period, and recalling that $E_A = r_e(x = 0)$, Eq. (32) gives:

$$E_A = D_{tot}(E_A) - w_{inh}\kappa E_A + I_A. \tag{40}$$

Second, during the memory period, from Eq. (33), the memory amplitude M_A is given by the steady-state activity of the neuron at the peak of the activity bump:

$$M_A = D_{tot}(E_A) - w_{inh}\kappa M_A. \tag{41}$$

Substituting the expression for $D_{tot}(E_A)$ from Eq. (41) into Eq. (40) and rearranging then gives:

$$M_A = E_A - \frac{I_A}{\gamma}, \tag{42}$$

where $\gamma = 1 + w_{inh}\kappa$. This relationship is analogous to Eq. (21) for the autapse model.

Substituting Eq. (42) into Eq. (39) then gives a relationship for the weight function in terms of M_A:

$$w(z_{max})R(z_{max}) = \frac{T_u}{M_A + \frac{I_A}{\gamma}}. \tag{43}$$

From Eqs. (41) and (38), the memory amplitude M_A is related to the number of activated dendrites by

$$\gamma M_A = 2\beta z_{max}. \tag{44}$$

Substituting this equation into Eq. (43) and using that $I_A = f^{-1}(M_A)$ gives:

$$w(z_{max})R(z_{max}) = \frac{\gamma T_u}{2\beta z_{max} + f^{-1}(\frac{2\beta z_{max}}{\gamma})}. \tag{45}$$

Since the above equation is a requirement on w(x) for any distance $x = \pm z_{max}$ from the peak of the bump that could occur with different stimulus amplitudes I_A, we finally have the general relation:

$$w(|x|) = \frac{\gamma T_u}{R(|x|)[2\beta|x| + f^{-1}(\frac{2\beta|x|}{\gamma})]}. \tag{46}$$

This equation is analogous to Eq. (25) for the autapse model, but modified by a factor $\gamma = 1 + w_{inh}\kappa$ that goes to one in the absence of inhibition, a spatial profile term $R(|x|)$, and a factor of two that accounts for the ring model weight profile being defined in terms of negative and positive x values.

The relations above can be rearranged to determine the relationship f() between the stimulus amplitude and memory amplitude for a given weight function w(x). Rearranging Eq. (43) gives:

$$I_A = \gamma[\frac{T_u}{w(z_{max})R(z_{max})} - M_A]. \tag{47}$$

Substituting the expression for z_{max} from Eq. (44) into this relation gives:

$$I_A = \gamma[\frac{T_u}{w(\frac{\gamma}{2\beta}M_A)R(\frac{\gamma}{2\beta}M_A)} - M_A] = f^{-1}(M_A), \tag{48}$$

which can be inverted to give $f(M_A)$. This formula parallels the analogous relation for the autapse case, Eq. (27). In the absence of recurrent inhibition, $\gamma = 1$, this expression only differs from the autapse case by the spatial profile term $R(x)$.

Linear memory-stimulus relationship. For the particular case of a linear relationship,

$$M_A = c(I_A - I_{A,th}), \tag{49}$$

between the amplitude of the external stimulus and the amplitude of the memory period activity bump illustrated in Figure 4G, we have:

$$I_A = f^{-1}(M_A) = \frac{M_A}{c} + I_{A,th}. \tag{50}$$

Substituting this relation into Eq. (46) gives:

$$w(x) = \frac{\gamma T_u}{R(|x|)[2\beta|x|(1 + \frac{1}{c\gamma}) + I_{A,th}]}. \tag{51}$$

Thus, analogous to the autapse case, the effective weight function $W(x) = w(x)R(x)$ to produce an (approximately) linear input-output function has the form $W(x) = \frac{a}{|x|+b}$. Conversely, for this form of weight function, the slope and threshold of the memory amplitude - stimulus amplitude relationships are given from Eqs. (43), (44), and (50) by

$$c = \frac{1}{\gamma(\frac{T_u\gamma}{2\beta a} - 1)} \text{ and } I_{A,th} = \frac{\gamma b T_u}{a}. \tag{52}$$

We note that, for a flat stimulus shape, $R(x) =$ constant, this reduces to the formula found for the rate-based autapse case. In addition, note that, although $w(x)R(x)$ decreases monotonically for monotonically decreasing stimulus profiles, $w(x)$ from Eq. (51) may not be monotonically decreasing in $|x|$ or may even blow up if $R(x)$ approaches zero at the edge of the bump. To offset this, and ensure that $w(x)$ is monotonically decreasing as $|x|$ increases, in all simulations, we manually set $w(x)$ to zero once it stopped monotonically decreasing.

Acknowledgments

This work was supported by the Air Force Office of Scientific Research (AFOSR) award number FA9550-22-1-0532 (JX, DLC) and NIH grants U19 NS132720 and R01 EY027036 (MSG). We thank the Goldman laboratory for helpful discussions and suggestions, and J. Bhasin for critical feedback on the manuscript.

References

[1] H Sebastian Seung, Daniel D Lee, Ben Y Reis, and David W Tank. The autapse: a simple illustration of short-term analog memory storage by tuned synaptic feedback. *Journal of computational neuroscience*, 9:171–185, 2000.

[2] Mark S Goldman, Albert Compte, and Xiao-Jing Wang. Neural integrator models. *Encyclopedia of neuroscience*, pages 165–178, 2009.

[3] Carlos D Brody, Ranulfo Romo, and Adam Kepecs. Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations. *Current opinion in neurobiology*, 13(2):204–211, 2003.

[4] Alfonso Renart, Pengcheng Song, and Xiao-Jing Wang. Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. *Neuron*, 38(3):473–485, 2003.

[5] Rishidev Chaudhuri and Ila Fiete. Computational principles of memory. *Nature neuroscience*, 19(3):394–403, 2016.

[6] Christos Constantinidis and Xiao-Jing Wang. A neural circuit basis for spatial working memory. *The Neuroscientist*, 10(6):553–565, 2004.

[7] Shun-ichi Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological cybernetics*, 27(2):77–87, 1977.

[8] Albert Compte, Nicolas Brunel, Patricia S Goldman-Rakic, and Xiao-Jing Wang. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral cortex*, 10(9):910–923, 2000.

[9] Rani Ben-Yishai, R Lev Bar-Or, and Haim Sompolinsky. Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences*, 92(9):3844–3848, 1995.

[10] Matthew F Panichello, Brian DePasquale, Jonathan W Pillow, and Timothy J Buschman. Error-correcting dynamics in visual working memory. *Nature communications*, 10(1):3366, 2019.

[11] Sung Soo Kim, Hervé Rouault, Shaul Druckmann, and Vivek Jayaraman. Ring attractor dynamics in the drosophila central brain. *Science*, 356(6340):849–853, 2017.

[12] Xiao-Jing Wang. 50 years of mnemonic persistent activity: quo vadis? *Trends in Neurosciences*, 44(11):888–902, 2021.

[13] Alexei A Koulakov, Sridhar Raghavachari, Adam Kepecs, and John E Lisman. Model for a robust neural integrator. *Nature neuroscience*, 5(8):775–782, 2002.

[14] Christos Constantinidis, Matthew N Franowicz, and Patricia S Goldman-Rakic. The sensory nature of mnemonic representation in the primate prefrontal cortex. *Nature neuroscience*, 4(3):311–316, 2001.

[15] Carlos D Brody, Adrián Hernández, Antonio Zainos, and Ranulfo Romo. Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. *Cerebral cortex*, 13(11):1196–1207, 2003.

[16] Ashwin Vishwanathan, Alex Sood, Jingpeng Wu, Alexandro D Ramirez, Runzhe Yang, Nico Kemnitz, Dodam Ih, Nicholas Turner, Kisuk Lee, Ignacio Tartavull, et al. Predicting modular functions and neural coding of behavior from a synaptic wiring diagram. *Nature Neuroscience*, pages 1–12, 2024.

[17] Philip L Smith and Roger Ratcliff. Psychology and neurobiology of simple decisions. *Trends in neurosciences*, 27(3):161–168, 2004.

[18] Joshua I Gold and Michael N Shadlen. The neural basis of decision making. *Annu. Rev. Neurosci.*, 30(1):535–574, 2007.

[19] Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438, 2006.

[20] Jeffrey M Beck, Wei Ji Ma, Roozbeh Kiani, Tim Hanks, Anne K Churchland, Jamie Roitman, Michael N Shadlen, Peter E Latham, and Alexandre Pouget. Probabilistic population codes for bayesian decision making. *Neuron*, 60(6):1142–1152, 2008.

[21] Ran Darshan and Alexander Rivkind. Learning to represent continuous variables in heterogeneous neural networks. *Cell Reports*, 39(1), 2022.

[22] David G Clark, LF Abbott, and Haim Sompolinsky. Symmetries and continuous attractors in disordered neural circuits. *bioRxiv*, pages 2025–01, 2025.

[23] Sukbin Lim and Mark S Goldman. Balanced cortical microcircuitry for spatial working memory based on corrective feedback control. *Journal of Neuroscience*, 34(20):6790–6806, 2014.

[24] Sam Carroll, Krešimir Josić, and Zachary P Kilpatrick. Encoding certainty in bump attractors. *Journal of computational neuroscience*, 37(1):29–48, 2014.

[25] Weronika Wojtak, Stephen Coombes, Daniele Avitabile, Estela Bicho, and Wolfram Erlhagen. A dynamic neural field model of continuous input integration. *Biological cybernetics*, 115(5):451–471, 2021.

[26] Anna Kutschireiter, Melanie A Basnak, Rachel I Wilson, and Jan Drugowitsch. Bayesian inference in ring attractor networks. *Proceedings of the National Academy of Sciences*, 120(9):e2210622120, 2023.

[27] Kathleen P Champion, Olivia Gozel, Benjamin S Lankow, G Bard Ermentrout, and Mark S Goldman. An oscillatory mechanism for multi-level storage in short-term memory. *Communications biology*, 6(1):829, 2023.

[28] Mark S Goldman, Joseph H Levine, Guy Major, David W Tank, and H Sebastian Seung. Robust persistent neural activity in a model integrator with multiple hysteretic dendrites per neuron. *Cerebral cortex*, 13(11):1185–1195, 2003.

[29] Maxim Nikitchenko and Alexei Koulakov. Neural integrator: A sandpile model. *Neural computation*, 20(10):2379–2417, 2008.

[30] Marcelo Camperi and Xiao-Jing Wang. A model of visuospatial working memory in prefrontal cortex: recurrent network and cellular bistability. *Journal of computational neuroscience*, 5(4):383–405, 1998.

[31] Christopher P Fall, Timothy J Lewis, and John Rinzel. Background-activity-dependent properties of a network model for working memory that incorporates cellular bistability. *Biological cybernetics*, 93(2):109–118, 2005.

[32] J Hounsgaard and O Kiehn. Serotonin-induced bistability of turtle motoneurones caused by a nifedipine-sensitive calcium plateau potential. *The Journal of physiology*, 414(1):265–282, 1989.

[33] RH Lee and CJ Heckman. Bistability in spinal motoneurons in vivo: systematic variations in rhythmic firing patterns. *Journal of neurophysiology*, 80(2):572–582, 1998.

[34] Eve Marder, LF Abbott, Gina G Turrigiano, Zheng Liu, and Jorge Golowasch. Memory from the dynamics of intrinsic membrane currents. *Proceedings of the national academy of sciences*, 93(24):13481–13486, 1996.

[35] Eugene M Izhikevich. Simple model of spiking neurons. *IEEE Transactions on neural networks*, 14(6):1569–1572, 2003.

[36] Erik Fransén, Babak Tahvildari, Alexei V Egorov, Michael E Hasselmo, and Angel A Alonso. Mechanism of graded persistent cellular activity of entorhinal cortex layer v neurons. *Neuron*, 49(5):735–746, 2006.

[37] Michael E Hasselmo and Chantal E Stern. Mechanisms underlying working memory for novel information. *Trends in cognitive sciences*, 10(11):487–493, 2006.

[38] Guy Major and David Tank. Persistent neural activity: prevalence and mechanisms. *Current opinion in neurobiology*, 14(6):675–684, 2004.

[39] Joel Zylberberg and Ben W Strowbridge. Mechanisms of persistent activity in cortical circuits: possible neural substrates for working memory. *Annual review of neuroscience*, 40:603–627, 2017.

[40] Guillaume Rodriguez, Matthieu Sarazin, Alexandra Clemente, Stephanie Holden, Jeanne T Paz, and Bruno Delord. Conditional bistability, a generic cellular mnemonic

mechanism for robust and flexible working memory computations. *Journal of Neuroscience*, 38(22):5209–5219, 2018.

[41] John E Lisman, Jean-Marc Fellous, and Xiao-Jing Wang. A role for nmda-receptor channels in working memory. *Nature neuroscience*, 1(4):273–275, 1998.

[42] Patrick A Shoemaker. Neural bistability and amplification mediated by nmda receptors: analysis of stationary equations. *Neurocomputing*, 74(17):3058–3071, 2011.

[43] Honi Sanders, Michiel Berends, Guy Major, Mark S Goldman, and John E Lisman. Nmda and gabab (kir) conductances: the "perfect couple" for bistability. *Journal of Neuroscience*, 33(2):424–429, 2013.

[44] Goldman Lab Github. Python scripts for the current paper.

[45] Si Wu, Kosuke Hamaguchi, and Shun-ichi Amari. Dynamics and computation of continuous attractors. *Neural computation*, 20(4):994–1025, 2008.

[46] Yoram Burak and Ila R Fiete. Fundamental limits on persistent activity in networks of noisy neurons. *Proceedings of the National Academy of Sciences*, 109(43):17645–17650, 2012.

[47] Sukbin Lim and Mark S Goldman. Noise tolerance of attractor and feedforward memory models. *Neural computation*, 24(2):332–390, 2012.

[48] Surya Ganguli, Dongsung Huh, and Haim Sompolinsky. Memory traces in dynamical systems. *Proceedings of the national academy of sciences*, 105(48):18970–18975, 2008.

[49] Dong-Sheng Wei, Yan-Ai Mei, Ashish Bagal, Joseph PY Kao, Scott M Thompson, and Cha-Min Tang. Compartmentalized and binary behavior of terminal dendrites in hippocampal pyramidal neurons. *Science*, 293(5538):2272–2275, 2001.

[50] Min Wang, Yang Yang, Ching-Jung Wang, Nao J Gamo, Lu E Jin, James A Mazer, John H Morrison, Xiao-Jing Wang, and Amy FT Arnsten. Nmda receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. *Neuron*, 77(4):736–749, 2013.

[51] Bram van Vugt, Timo van Kerkoerle, Devavrat Vartak, and Pieter R Roelfsema. The contribution of ampa and nmda receptors to persistent firing in the dorsolateral prefrontal cortex in working memory. *Journal of Neuroscience*, 40(12):2458–2470, 2020.

[52] Shengtao Yang, Hyojung Seo, Min Wang, and Amy FT Arnsten. Nmdar neurotransmission needed for persistent neuronal firing: potential roles in mental disorders. *Frontiers in Psychiatry*, 12:337, 2021.

[53] Xiao-Jing Wang. Synaptic basis of cortical persistent activity: the importance of nmda receptors to working memory. *Journal of Neuroscience*, 19(21):9587–9603, 1999.

[54] Jackie Schiller, Guy Major, Helmut J Koester, and Yitzhak Schiller. Nmda spikes in basal dendrites of cortical pyramidal neurons. *Nature*, 404(6775):285–289, 2000.

[55] Guy Major, Alon Polsky, Winfried Denk, Jackie Schiller, and David W Tank. Spatiotemporally graded nmda spike/plateau potentials in basal dendrites of neocortical pyramidal neurons. *Journal of neurophysiology*, 99(5):2584–2601, 2008.

[56] Christof Koch, Tomaso Poggio, and Vincent Torre. Retinal ganglion cells: a functional interpretation of dendritic morphology. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 298(1090):227–263, 1982.

[57] Panayiota Poirazi, Terrence Brannon, and Bartlett W Mel. Pyramidal neuron as two-layer neural network. *Neuron*, 37(6):989–999, 2003.

[58] Panayiota Poirazi, Terrence Brannon, and Bartlett W Mel. Arithmetic of subthreshold synaptic summation in a model ca1 pyramidal cell. *Neuron*, 37(6):977–987, 2003.

[59] Panayiota Poirazi and Athanasia Papoutsi. Illuminating dendritic function with computational models. *Nature reviews neuroscience*, 21(6):303–321, 2020.

[60] Alon Polsky, Bartlett W Mel, and Jackie Schiller. Computational subunits in thin dendrites of pyramidal cells. *Nature neuroscience*, 7(6):621–627, 2004.

[61] Greg J Stuart and Nelson Spruston. Dendritic integration: 60 years of progress. *Nature neuroscience*, 18(12):1713–1721, 2015.

[62] Victor Hugo Cornejo, Netanel Ofer, and Rafael Yuste. Voltage compartmentalization in dendritic spines in vivo. *Science*, 375(6576):82–86, 2022.

[63] Roberto Araya. Dendritic morphology and function. *Neuroscience in the 21st Century: From Basic to Clinical*, pages 571–606, 2022.

[64] Nicolas Vecoven, Damien Ernst, and Guillaume Drion. A bio-inspired bistable recurrent cell allows for long-lasting memory. *Plos one*, 16(6):e0252676, 2021.

[65] Jonathan Green, Atsuko Adachi, Kunal K Shah, Jonathan D Hirokawa, Pablo S Magani, and Gaby Maimon. A neural circuit architecture for angular integration in drosophila. *Nature*, 546(7656):101–106, 2017.

[66] Daniel Turner-Evans, Stephanie Wegener, Hervé Rouault, Romain Franconville, Tanya Wolff, Johannes D Seelig, Shaul Druckmann, and Vivek Jayaraman. Angular velocity integration in a fly heading circuit. *Elife*, 6:e23496, 2017.

[67] Matteo Carandini and David J Heeger. Normalization as a canonical neural computation. *Nature reviews neuroscience*, 13(1):51–62, 2012.

[68] Jayeeta Basu, Jeffrey D Zaremba, Stephanie K Cheung, Frederick L Hitti, Boris V Zemelman, Attila Losonczy, and Steven A Siegelbaum. Gating of hippocampal activity, plasticity, and memory by entorhinal cortex long-range inhibition. *Science*, 351(6269):aaa5694, 2016.

[69] Olesia M Bilash, Spyridon Chavlis, Cara D Johnson, Panayiota Poirazi, and Jayeeta Basu. Lateral entorhinal cortex inputs modulate hippocampal dendritic excitability by recruiting a local disinhibitory microcircuit. *Cell reports*, 42(1), 2023.

[70] H Sebastian Seung and Haim Sompolinsky. Simple models for reading neuronal population codes. *Proceedings of the national academy of sciences*, 90(22):10749–10753, 1993.

[71] Alexandre Pouget, Peter Dayan, and Richard S Zemel. Inference and computation with population codes. *Annual review of neuroscience*, 26(1):381–410, 2003.

[72] Alexandre Pouget, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains: knowns and unknowns. *Nature neuroscience*, 16(9):1170–1178, 2013.

[73] Robert Wilson and Leif Finkel. A neural implementation of the kalman filter. *Advances in neural information processing systems*, 22, 2009.