Supplemental Figures

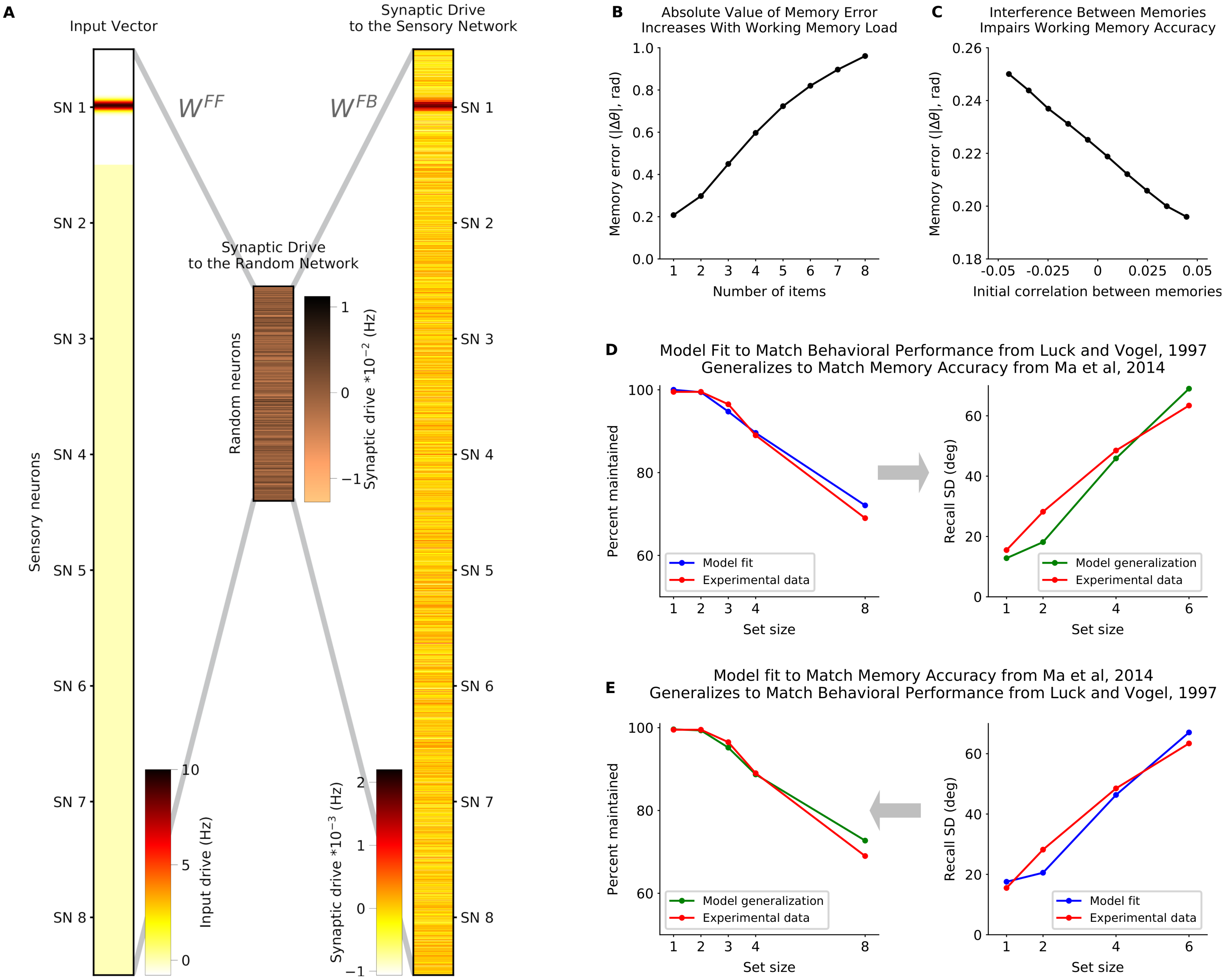

Figure S1 – **Random connections allow for flexibility in memories but limit capacity.** Caption next page.

Figure S1 – **Random connections allow for flexibility in memories but limit capacity.** Related to Figures 1, 3 and 5. **(A)** The synaptic drive fed back into a sensory sub-network from the random network closely matches its own representation. The left column shows the synaptic drive for the entire sensory network in response to a sensory input into sensory sub-network 1. The middle column shows the synaptic drive to the random network resulting from this sensory input (estimated by passing synaptic drive from sensory network through W^{FF}). The right column shows the feedback synaptic drive to the sensory network from the random network (estimated by passing the synaptic drive of the random network (middle) through W^{FB}). The similarity between the sensory input (left) and feedback input (right) sustains memory representations. **(B-C)** Absolute value of circular error increases with memory load and with interference between memories. **(B)** Absolute value of the circular error computed from ML spike decoding after 1 second of network simulation, as a function of load. Decoding time window is 100msec. **(C)** Absolute value of circular error as a function of the correlation between two inputs in two sensory sub-networks. Only simulations were both memories are maintained were considered (i.e. the error is not due to forgetting). **(D-E)** Model parameters can be quantitatively fit to match behavioral results. **(D)** Model parameters were fit to match human behavioral performance on a change detection task [Luck and Vogel, 1997] (left). The memory accuracy of the resulting model matches experimental observations of memory accuracy from [Ma et al., 2014] (right). See Methods for details. **(E)** When model parameters were fit to match memory accuracy (right), memory performance of the network generalized to match experimental results (left).

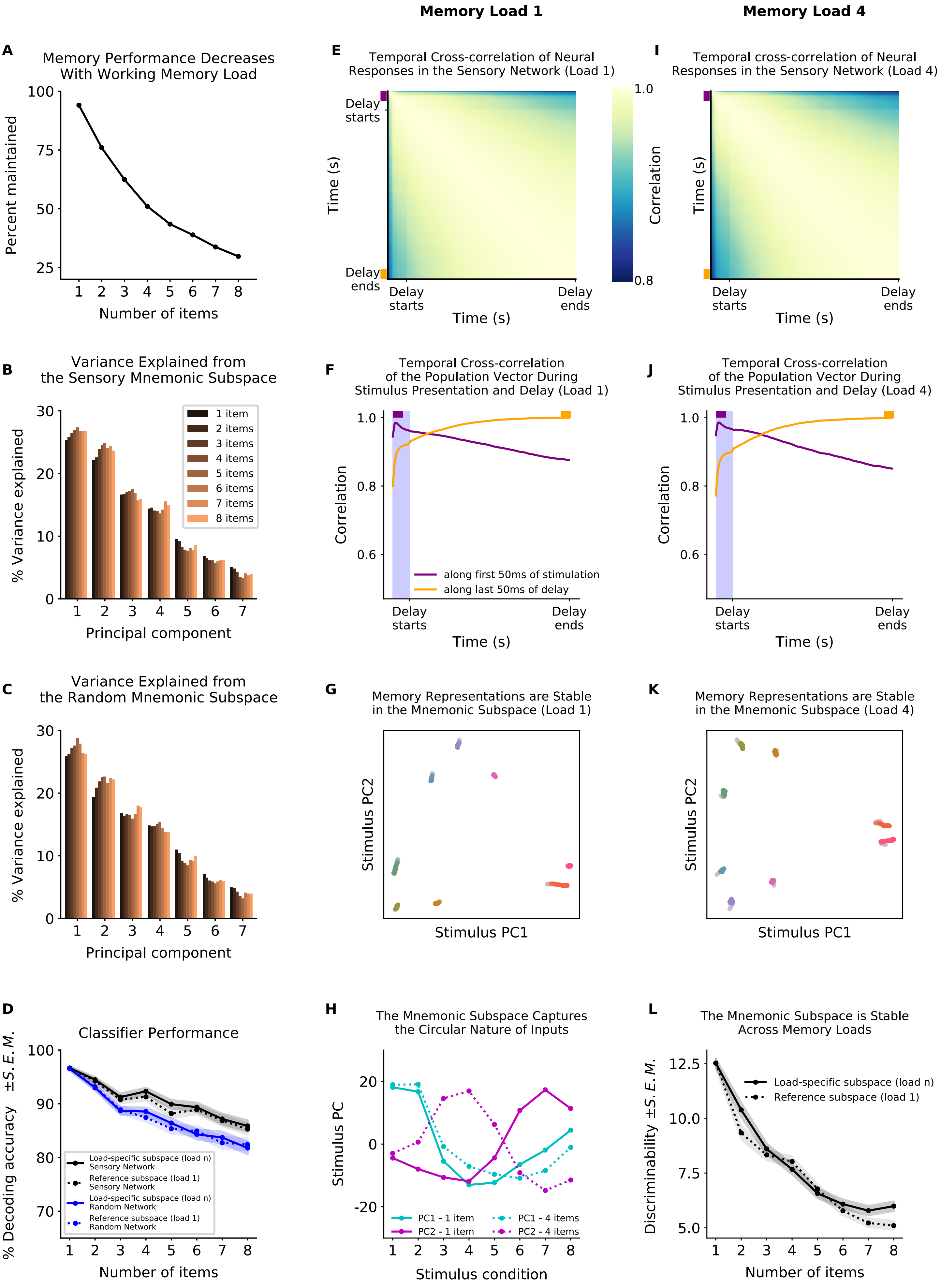

Figure S2 – **Dynamics of memory representations** Caption next page.

Figure S2 – **Dynamics of memory representations**. Related to Figure 6. **(A)** Adding direct sensory inputs into the random network and recurrence within the random network did not alter network performance. The percentage of correct memories as a function of load is similar to Fig. 3A. Network included direct transient input into the random network and weak recurrence within the random network (as in Fig. 6; see Methods for details). **(B-C)** The dimensionality of the mnemonic subspace in the **(B)** sensory network and **(C)** random network is similar across memory loads ('number of items'). This is reflected in the similar percentage of variance explained by each principal component across memory loads (principal components are estimated for time-averaged delay activity as in [Murray et al., 2016], see Methods for details). **(D)** Performance of a centroid classifier was similar when the mnemonic subspace was defined for the current load (solid lines) or when the mnemonic subspace was defined for a single item (load 1). Shown for both sensory network and random network. Decodability was not significantly different across subspaces (a two-sample Wald test for load-specific subspace versus reference subspace gives $p > 0.1$ for all loads). **(E-L)** Dynamics in sensory network. Follows the structure of Figure 6 which shows dynamics in the random network. **(E)** Temporal cross-correlation of neural activity in the sensory network shows activity during the stimulation is more correlated with activity during the delay period than for the random network (Fig. 6A). **(F)** Slices of the matrix represented in **E**: correlation of population state from the first 50ms of the stimulus period (purple) and the last 50ms of the delay period (orange) against all other times. **(G)** As in the random network, the memory is stable in the sensory network. Here the response of the sensory network population is projected onto the mnemonic subspace (defined by the first two principal components of time-averaged activity, see Methods for details). Each trace corresponds to the response to a different input into sensory sub-network 1, shown over time (from lighter to darker colors). **(H)** Mnemonic subspace is defined by two orthogonal, quasi-sinusoidal representations of inputs, capturing the circular nature of sensory sub-networks. These representations were similar between load 1 and load 4, with only an inversion of PC2. **(I-K)** Same as **E-G** but for a load of 4. Only includes simulations for which the memory in sensory sub-network 1 was maintained (other three memories can be forgotten). **(L)** The mnemonic subspace is stable across working memory load. Decodability of memory was measured as discriminability, d', between inputs (see Methods for details and **D** for centroid classifier). Decodability is similar when mnemonic subspace was defined for a single input (dashed line) or specific for each load (solid line). Decodability was not significantly different across subspaces for load 3, 4, and 5 (a two-sample Wald test for optimized subspace versus reference subspace gives respectively $p = 0.24$, $p = 0.55$, and $p = 0.44$). For load 2, 6, 7, and 8, the two-sample Wald test gives $p = 0.022$, $p = 0.0080$, $p = 10^{-4}$ and $p < 10^{-5}$ respectively. Decodability is reduced with load ($p < 0.001$).

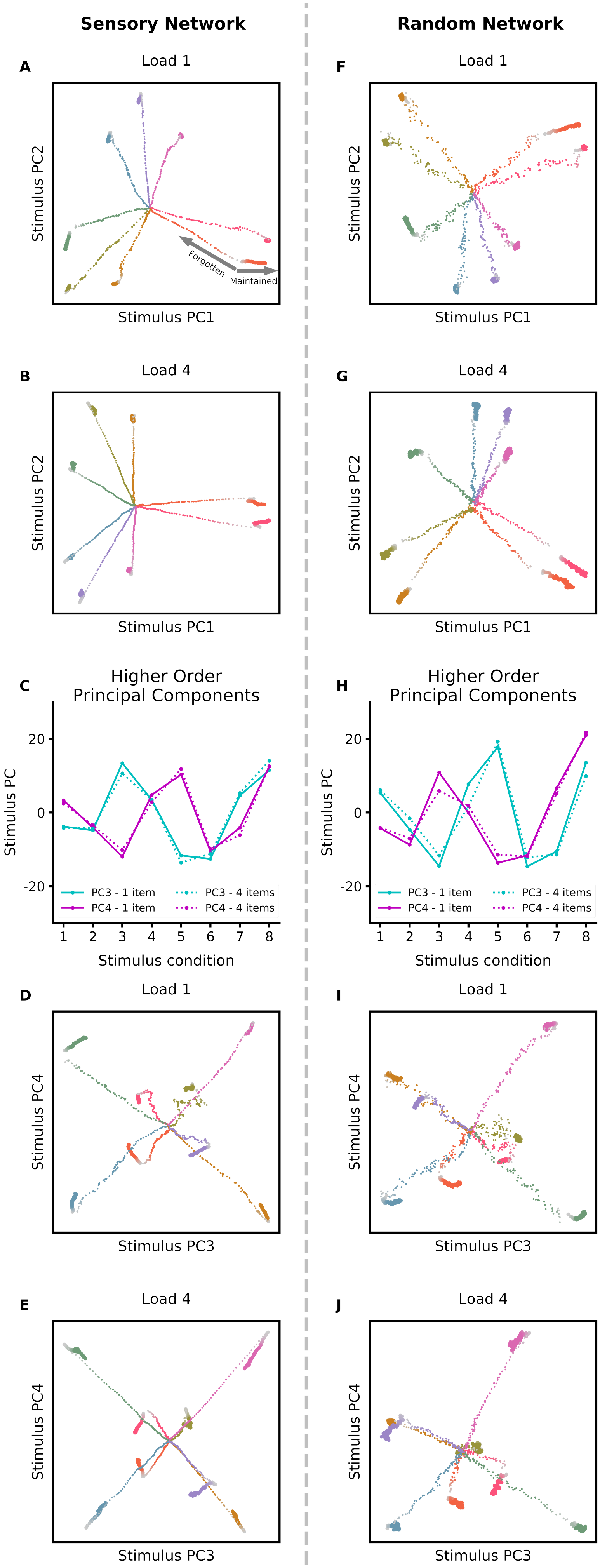

Figure S3 – **Memories fail by collapsing to a null state.** Caption next page.

Figure S3 – **Memories fail by collapsing to a null state.** Related to Figure 6 and S2. **(A-B)** Sensory network activity is projected onto the mnemonic subspace (defined by the first two principal components of time-averaged activity, see Methods for details). Each trace corresponds to the response to a different input into sensory sub-network 1, shown over time (from lighter to darker colors). Trials for when the memory is successfully maintained are shown in thick lines and trials for when the memory is forgotten are shown as thin lines. Note that memories fail by collapsing to a 'null' state with no activity. Timecourses are shown for a memory load of **(A)** 1 item and **(B)** 4 items. **(C)** The third and fourth principal components are shown across stimulus space These components are higher order harmonics of the circular space (see Figure S2H for PC1 and PC2). **(D-E)** The response of the sensory network population projected onto the subspace defined by PC3 and PC4 (as in **A** and **B**, see Methods for details). Note that memories fail by collapsing to a 'null' state with no activity. Timecourses are shown for a memory load of **(D)** 1 item and **(E)** 4 items. **(F-J)** As in **A-E** but for the response of neurons in the random network.

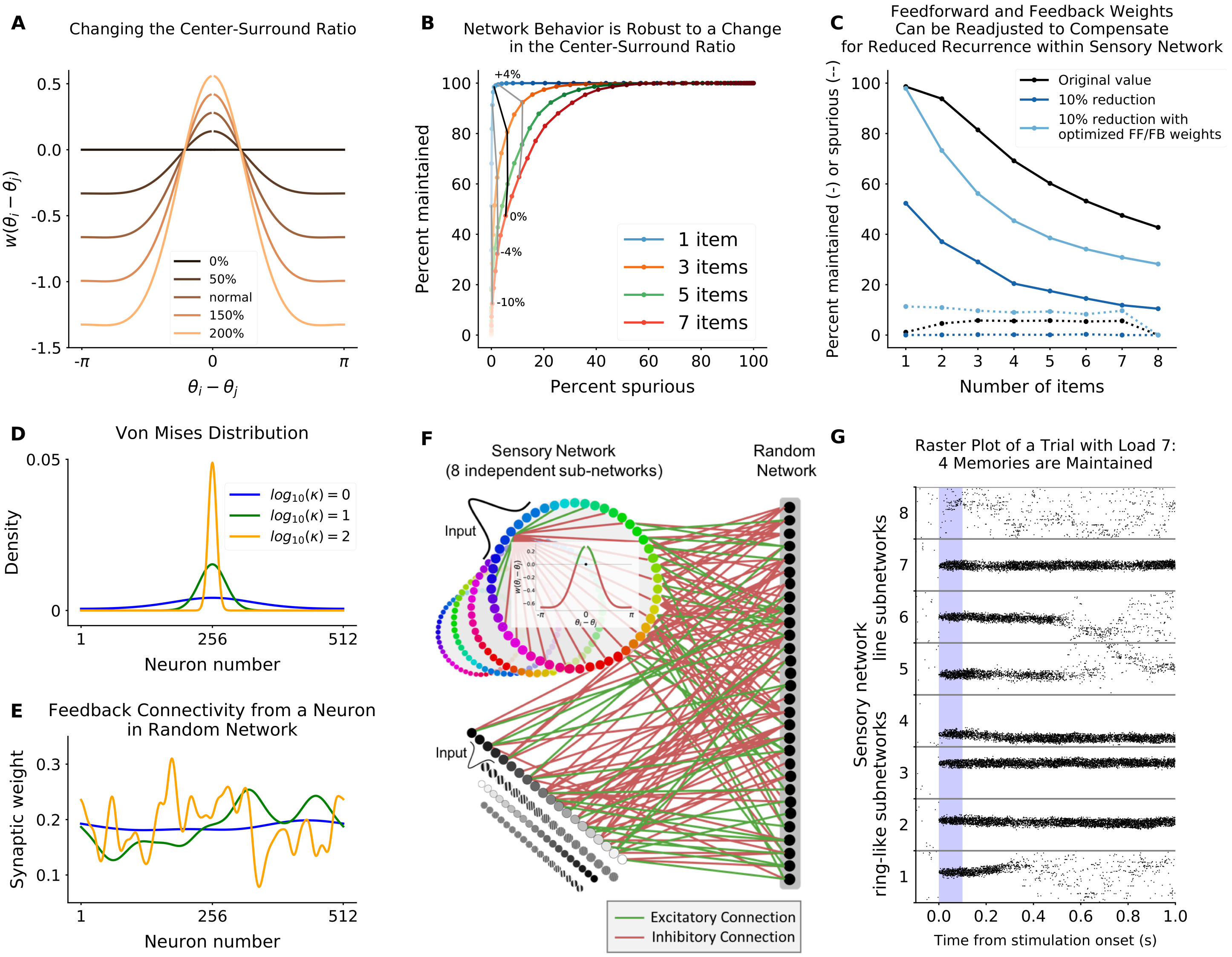

Figure S4 – Caption next page.

Figure S4 – Related to Figure 7. **(A-C) The network is robust to changes in the strength of recurrent interactions within the sensory sub-networks. (A)** The strength of the center-surround connectivity was changed by multiplying by a constant from 0% to 200% (see Methods). **(B)** ROC-like plot displaying the behavior of the network (i.e. percentage of maintaining a memory, versus percentage of creating a spurious) as the center-surround is modulated by a percentage from the original value, all other parameters being fixed. This is shown across memory loads 1, 3, 5 and 7 (colored lines). Similar to Figure 7A which shows network performance as feedforward/feedback weights are changed. **(C)** Memory performance is reduced when the center-surround architecture is weakened. Black lines show performance of original network; dark blue lines show performance of a network with a 10% reduction in center-surround strength within the sensory network. However, this reduction can be partially rescued by changing the feedforward/feedback weights between the sensory and the random networks (light blue), with all other parameters being fixed. **(D-E) Example Von Mises distributions for spreading of feedback connections.** In Fig. 7D, each feedback excitatory connection was distributed according to a Von Mises distribution with varying concentrations. **(D)** Example Von Mises distributions for different concentration parameters. **(E)** An example of feedback connectivity as a sum of VonMises from a neuron from the random network to a sensory sub-network (neuron 1 to 512) when $\kappa = 1$, $\kappa = 10$ and $\kappa = 100$ (colors as in **D**). **(F-G) Network performance did not depend on architecture of sensory sub-networks. (F)** Schematic of alternative model architecture where the sensory network is now composed of 4 ring-like sub-networks and 4 line sub-networks. All other parameters remained the same. **(G)** Raster plot of an example trial with 8 sensory sub-networks (512 neurons each) randomly connected to the same random network (1024 neurons). 7 sensory sub-networks (4 ring, 3 line sub-networks) receive a Gaussian input for 0.1 seconds during the 'stimulus presentation' period (shaded blue region). As for the original architecture, memories are maintained but the network has a capacity limit.

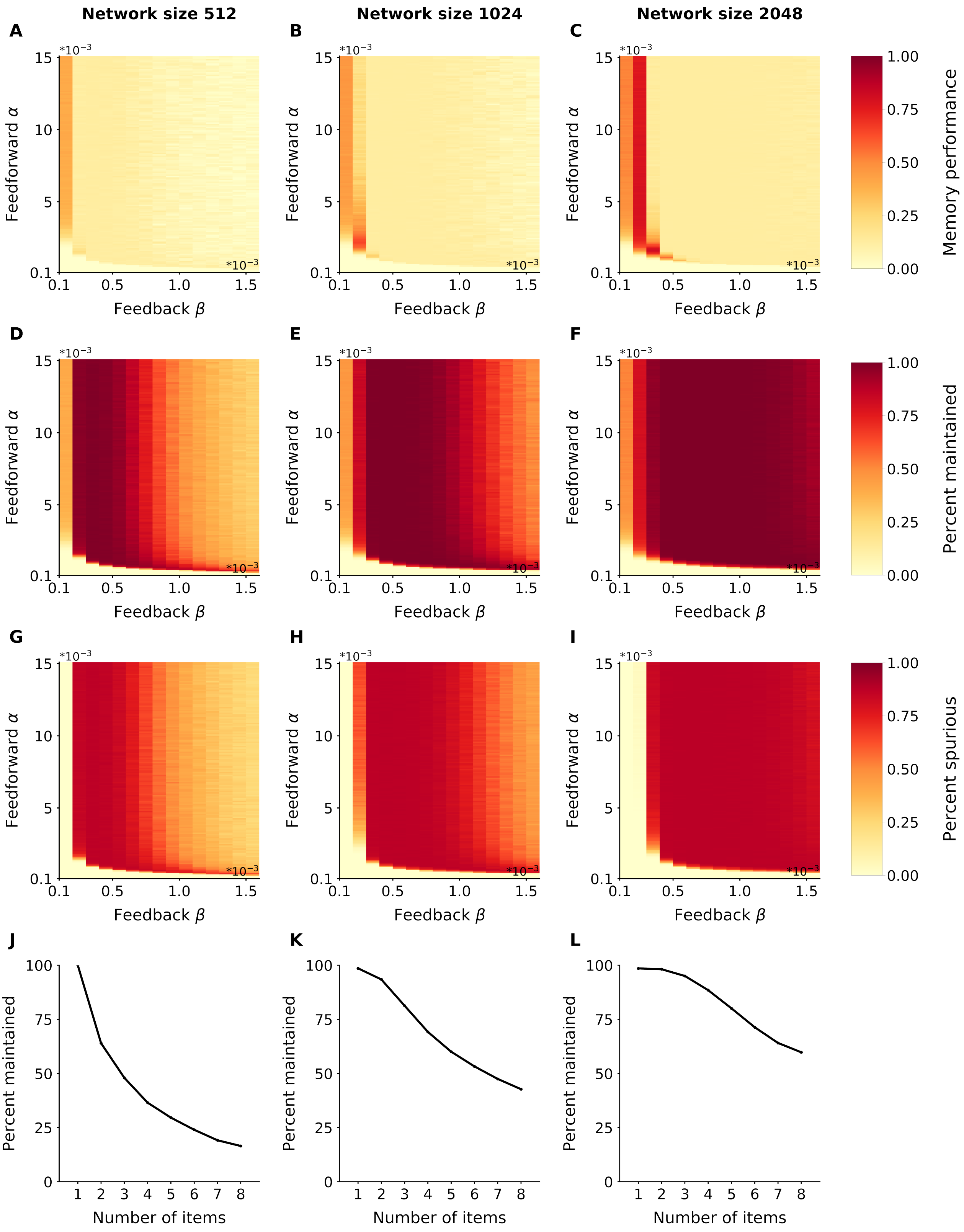

Figure S5 – **Increasing the size of the random network increases memory capacity.** Related to Figure 7. **(A-C)** Memory performance (color axis) as a function of feedforward and feedback weights, respectively for a random network size of **(A)** 512, **(B)** 1024, or **(C)** 2048. **(D-F)** Probability of maintaining a memory (color axis) as a function of feedforward and feedback weights, respectively for a random network size of 512, 1024, or 2048. **(G-I)** Probability of creating a spurious memory (color axis) as a function of feedforward and feedback weights, respectively for a random network size of 512, 1024, or 2048. **(J-L)** Percentage of correct memories, as a function of load, for the best set of parameters (α,β). Note the increase in capacity as the size of the random network increases, relative to the size of the sensory network.

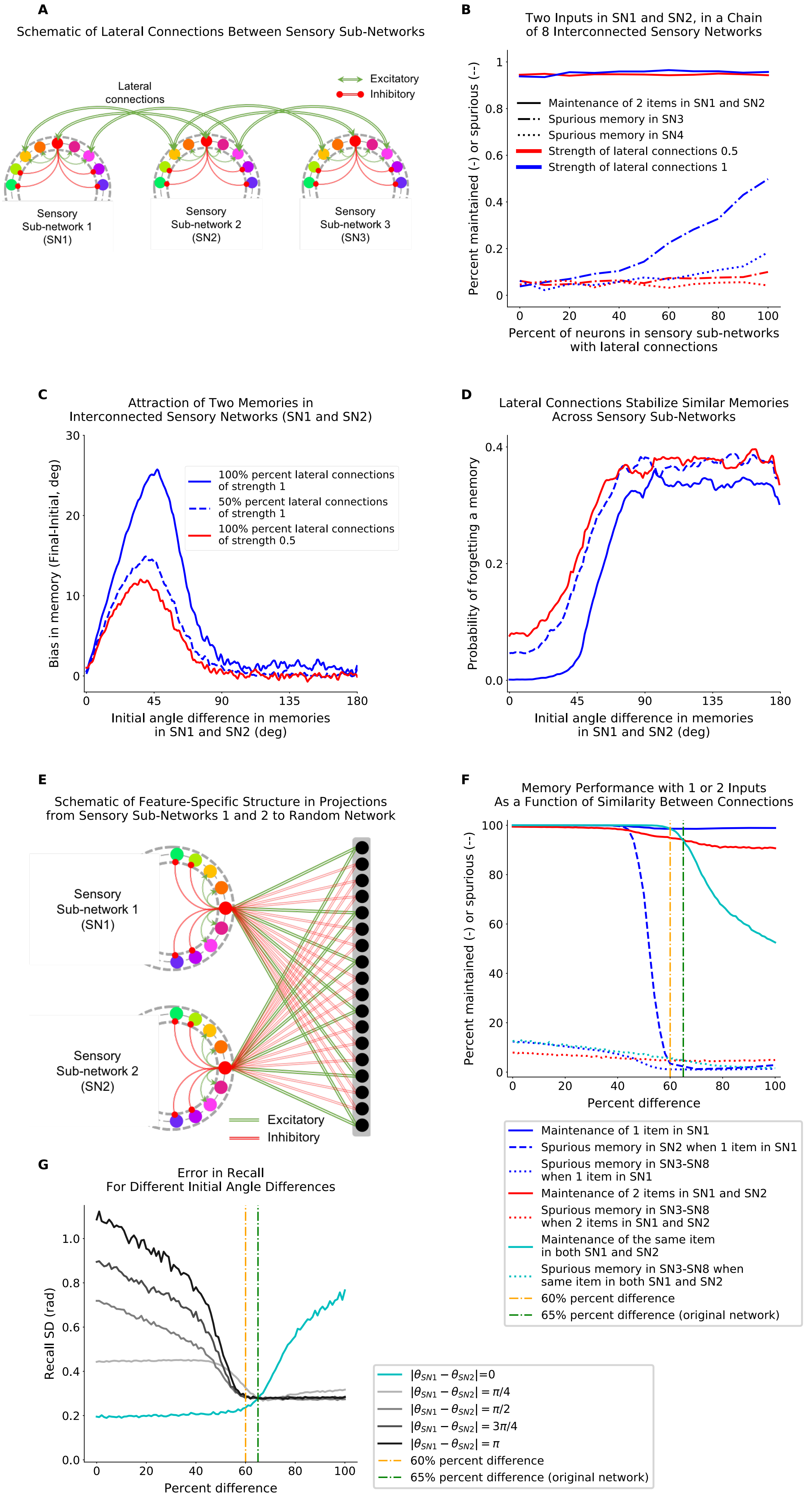

Figure S6 – Caption next page.

Figure S6 – Related to Figure 7. **(A-D) Lateral connection between ring-like sensory sub-networks causes interactions between memories. (A)** Schematic of lateral connections. The 8 ring-like sensory sub-networks (SN) are connected by excitatory connections to neighboring rings. Although only the connections between SNs 1, 2, and 3 are represented, all SNs were connected in a chain (see Methods). **(B)** Proportion of spurious memories created in sensory SNs 3 and 4 when both SN 1 and 2 received inputs. Spurious memories increased if the strength of the excitatory overlapping connections was increased (compare weight of 0.5, red line, to 1.0, blue line) or the percentage of neurons with lateral connections was increased (x-axis). **(C)** Bias of memories in SN1 and SN2 as a function of the initial angle between the two inputs. Bias is measured as the angular difference between the initial angle and final angle of the memory. By definition, positive bias reflects attraction. This effect is similar when the frequency or strength of lateral connections is varied (different lines). **(D)** Stabilization of memories in SN1 and SN2 as a function of the initial angle between the two inputs. The probability of forgetting a memory (y-axis) is reduced if the difference in inputs (x-axis) is smaller. This effect is similar when the frequency or strength of lateral connections is varied (line labels as in **C**). **(E-G) Testing the impact of overlap in the projections from sensory sub-networks to the random network. (E)** Schematic of the feature specific structure between sensory sub-network 1 (SN1) and sensory sub-network 2 (SN2). The 'percent difference' between the projections from SN1 and SN2 varied from 0 to 100% (see Methods for details). Schematic shows an example of 0% difference: the projection of the red neuron from SN1 to the random network is the same as the projection of the red neuron in SN2. All other pairs of neurons in SN1 and SN2 have similarly overlapping projections. **(F)** Increasing overlap (decreasing difference) improves maintenance when two inputs are presented (solid red line), but also results in the creation of a spurious memory when only one input is presented (dashed blue line). On the contrary, having two uncorrelated weight matrices (percent difference of 100) impairs maintenance of two identical initial inputs in SN1 and SN2 (solid cyan line). The range 60-65% is optimal; maintaining memories without creating spurious memories. Our initial model has a percent difference of $1-\gamma = 65\%$, in this optimal zone. **(G)** The error in recall when two inputs are presented to SN1 and SN2 depends both on the initial distance between inputs and on the overlap of projections.

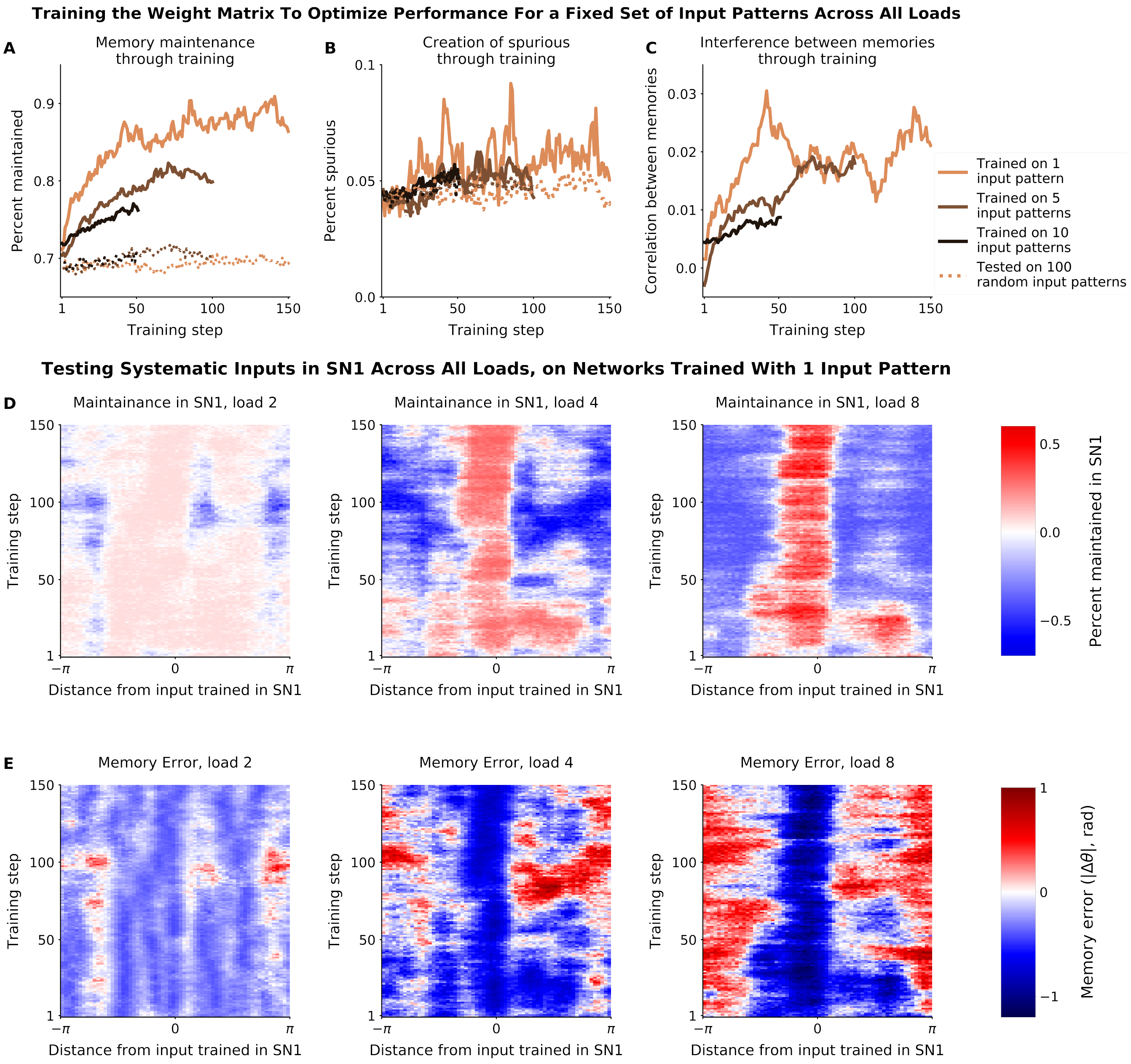

Figure S7 – **Training the weight matrix to optimize performance on fixed inputs does not generalize to new inputs.** Caption next page.

Figure S7 – **Training the weight matrix to optimize performance on fixed inputs does not generalize to new inputs.** Related to Figure 7E. As detailed in the Methods, a network was trained to maximize performance for 1, 5 or 10 input patterns across all loads (from 1 to 8). **(A-B)** Training improved the percent of maintained memories **(A)** and minimized the percent of spurious memories **(B)**. The difference between these statistics was used to compute memory performance in Figure 7E. As detailed in the main text, learning was slower when the number of inputs to be simultaneously optimized was increased. Dashed lines show network performance on a set of 100 random input patterns, across all loads. The optimization did not generalize beyond the set of trained inputs. **(C)** Memory performance was increased by increasing the correlation in the random network of trained memories. As seen in Figures S8 and S6F,G, too much correlation leads to spurious memories, likely leading to the plateau in correlation during training. **(D-E)** Training interfered with the maintenance of other memories. We tested how well other inputs were remembered in the network trained to remember 1 input pattern. Both memory performance **(D)** and accuracy **(E)** of the network were quantified across training steps (y-axis). We systematically varied the input into sensory sub-network 1 (SN1), relative to the trained input into SN1 (x-axis) while also varying the memory load (load 2, 4, and 8 are shown in the left, middle, and right columns, respectively). Load was increased by providing inputs to a random subset of sub-networks other than SN1. These sub-networks always received their trained input during testing. Both memory performance and accuracy for the trained input improved over training. However, training disrupted the ability of the network to successfully remember inputs different from the trained input.

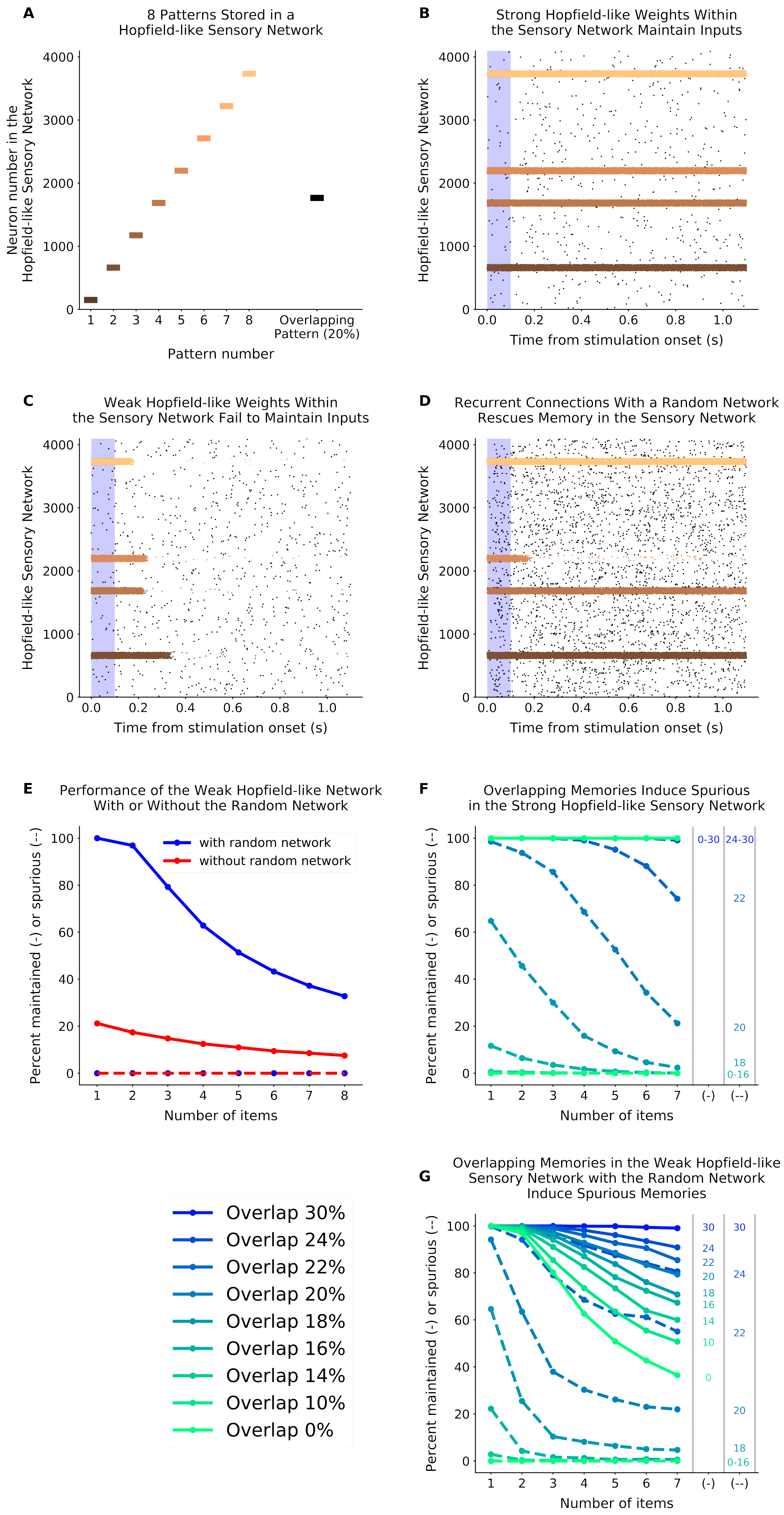

Figure S8 – **Replacing the ring-like sub-networks by a Hopfield-like sensory network to show the flexibility of the proposed architecture.** Caption next page.

Figure S8 – **Replacing the ring-like sub-networks by a Hopfield-like sensory network to show the flexibility of the proposed architecture.** Related to Figure 8. **(A)** Illustration of the eight 'Hopfield' patterns used in **B,C,D,E**, as well as the overlapping pattern used in **F,G**. **(B)** Raster plot of a simulated trial of the Hopfield-like sensory network alone, with 'stronger weights' ($\lambda = 600$, see Methods). All presented patterns are maintained throughout the delay. **(C)** Same simulation as **B**, but now the weights are depleted ($\lambda = 800$). No pattern is maintained. **(D)** Same simulation as **C**, but with the addition of the random network. Three patterns are maintained, out of the four presented initially. **(E)** Performance of the Hopfield-like sensory network with 'weaker weights' ($\lambda = 800$, see Methods), with or without the random network. As with the other sensory network architectures, representations interfere in the random network, leading to a decrease in performance with the number of items in memory. **(F-G)** One pattern is replaced by a new pattern that overlapped with pattern 4 (see **A**). 'Percent maintained' (solid lines) refers to the fraction of trials where the initial pattern 4 is maintained. 'Percent spurious' (dashed lines) refers to the fraction of trials when the new overlapping pattern was spuriously activated when pattern 4 was presented. The load is varied by adding, on top of pattern 4, other non-overlapping patterns (patterns 1-3 and 5-7 in **A**). **(F)** The Hopfield-like sensory network with stronger weights, and without the random network always maintains the initial pattern 4. However, it also creates a spurious memory of the overlapped pattern above 16%, with higher probability when the overall memory load is lower. **(G)** The random network is added to the Hopfield-like sensory network with weaker weights. The addition of an overlapping memory helps to stabilize memory pattern 4 (solid lines, increasing performance for increasing overlap). However, as in the Hopfield-like sensory network with strong weights, too much overlap can lead to spurious activation of the overlapping pattern (dashed lines).