

Supplementary Figures and Supplementary Methods for “Distinguishing examples while building concepts in hippocampal and artificial networks”

Louis Kang^{*1} and Taro Toyoizumi²

¹Neural Circuits and Computations Unit, RIKEN Center for Brain Science

²Laboratory for Neural Computation and Adaptation, RIKEN Center for Brain Science

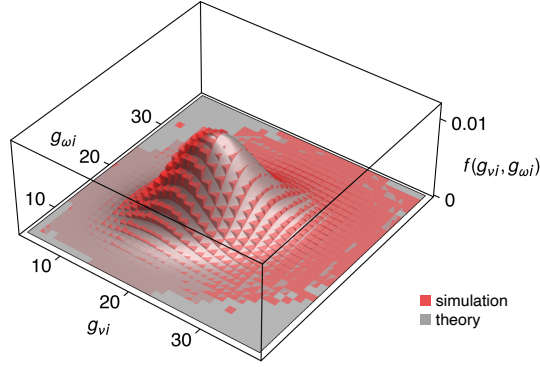
Contents

Supplementary Figures	2
Supplementary Figure 1	2
Supplementary Figure 2	2
Supplementary Figure 3	3
Supplementary Figure 4	4
Supplementary Figure 5	5
Supplementary Figure 6	7
Supplementary Figure 7	9
Supplementary Figure 8	11
Supplementary Methods	12
Decorrelation in binary feedforward networks	12
CA3 model with random binary patterns	18
CA3 model behavior during oscillating threshold	18
Experimental data preprocessing	19
Experimental data aggregate analysis	20
References	21

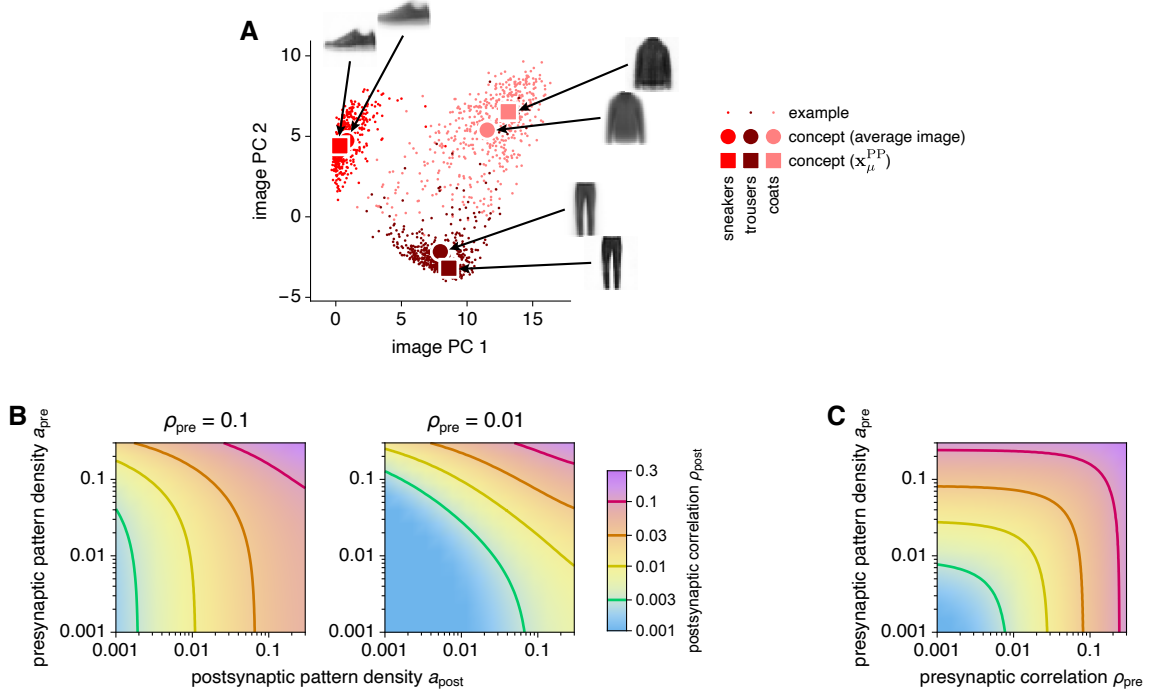
N.B.: All equation numbers below refer to this document and not the main text unless explicitly stated.

^{*}louis.kang@riken.jp

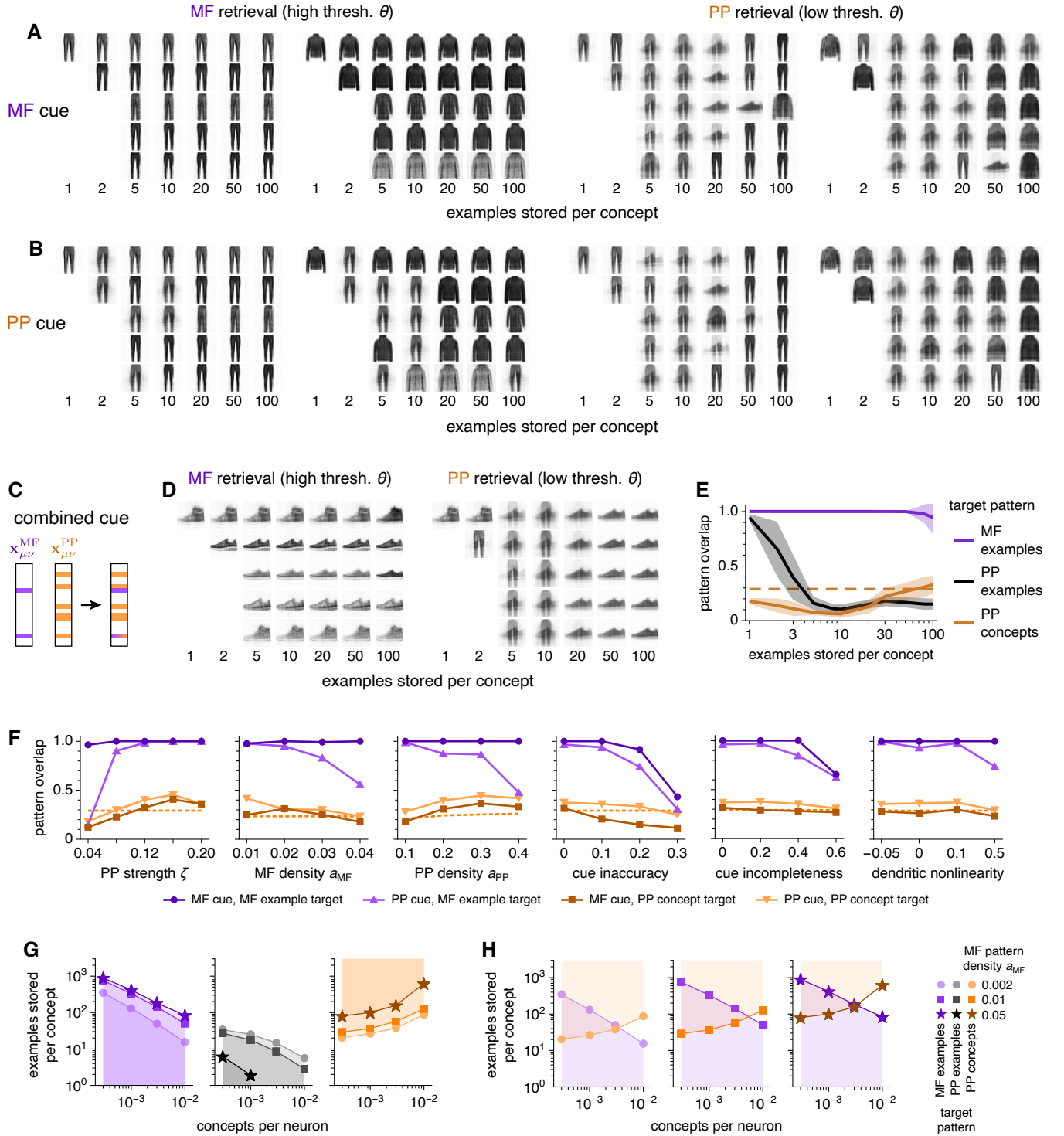
Supplementary Figures



Supplementary Figure 1: Extended results on decorrelation in binary feedforward networks. Joint probability distribution of total inputs $G_{\nu i}$ and $G_{\omega i}$ on postsynaptic neuron i for two patterns $\nu \neq \omega$ (Eq. 12). The theoretically derived probability density function $f(g_{\nu i}, g_{\omega i})$ (Eq. 25) agrees with simulation results. Each g is a sample of the corresponding random variable G . Simulation parameters are 1000 presynaptic neurons, 1000 postsynaptic neurons, synaptic connection probability 0.1, 100 random activity patterns, presynaptic pattern density 0.1, and presynaptic correlation 0.3.



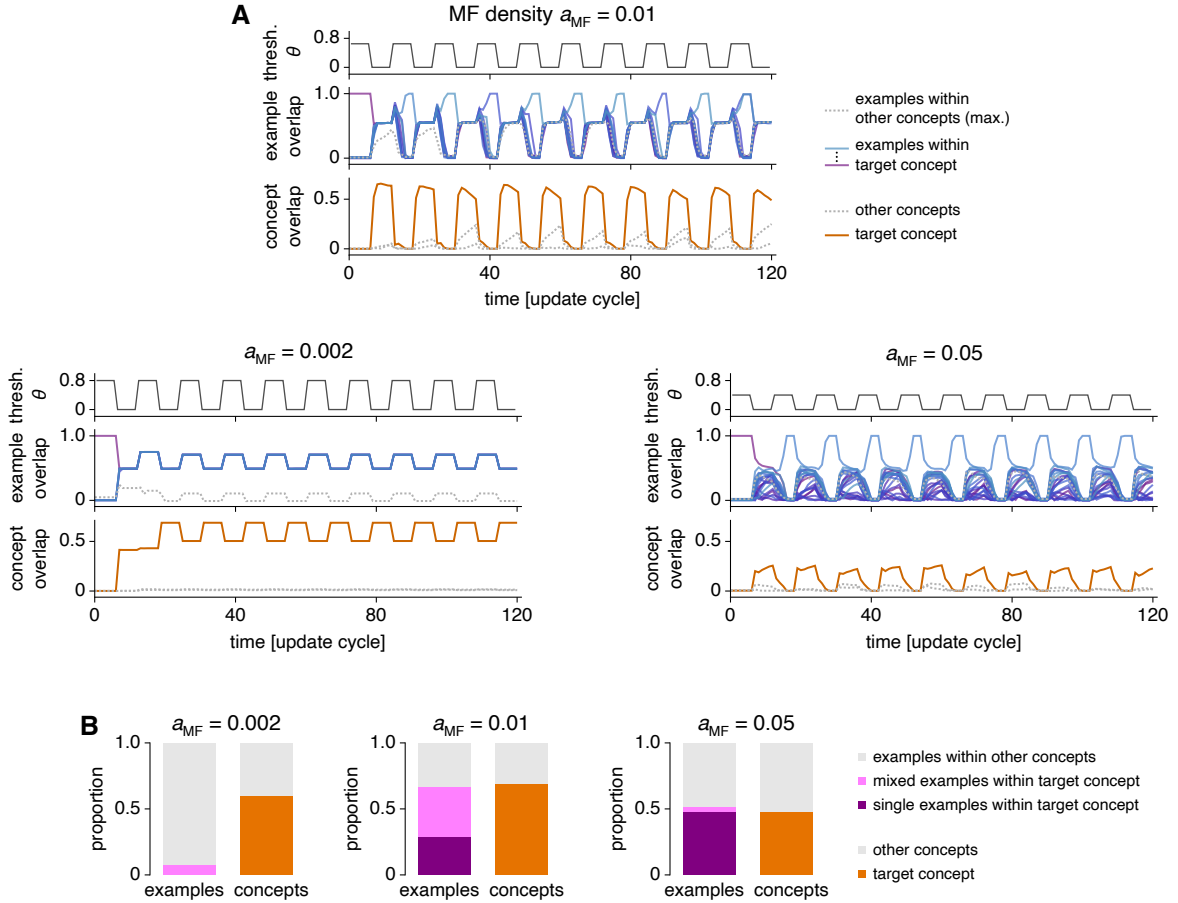
Supplementary Figure 2: Extended results for Fig. 2 of the main text. **(A)** Example images and two definitions of concepts projected along the first two principal components (PCs) of the example dataset. Each concept contains 512 examples. Averaged images are formed by averaging pixel intensities. PP concept target patterns $\mathbf{x}_{\mu}^{PP}$ are formed according to Eq. 13 of the main text and then passed through the visualization pathways in Fig. 2F of the main text. **(B, C)** Postsynaptic correlation ρ_{post} in binary feedforward networks in Fig. 2D, E. **(B)** ρ_{post} as a function of presynaptic pattern density a_{pre} and postsynaptic pattern density a_{post} for two values of presynaptic correlation ρ_{pre} . On the left, $\rho_{\text{pre}} = 0.1$ corresponds to the red line, and on the right, $\rho_{\text{pre}} = 0.01$ corresponds to the yellow line. Decorrelation occurs in the regions below and to the left of these lines. **(C)** ρ_{post} as a function of presynaptic pattern density a_{pre} and presynaptic correlation ρ_{pre} for postsynaptic pattern density $a_{\text{post}} = 0.1$. Source data are provided as a Source Data file.



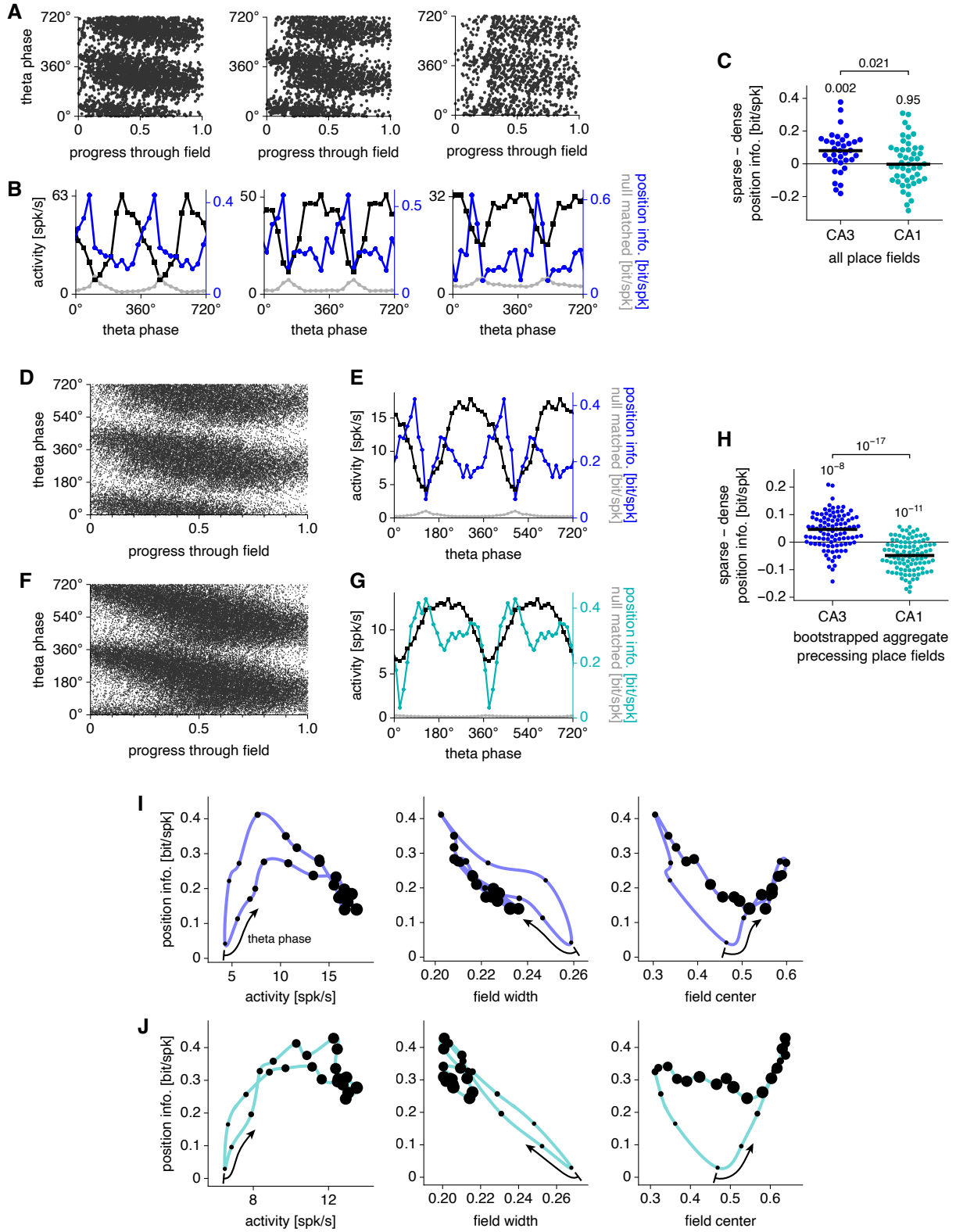
Supplementary Figure 3: Extended results for Fig. 3 of the main text. **(A)** Similar to Fig. 3D, but for trouser and coat concepts. **(B)** Similar to Fig. 3F, but for trouser and coat concepts. **(C–E)** Results for cues that combine active neurons from both MF and PP encodings. **(C)** Combined cues formed by the neuron-wise or operation. **(D)** Similar to Fig. 3D, F, but for combined cues. **(E)** Similar to Fig. 3E, G, but for combined cues. (Continued on the next page.)

Supplementary Figure 3: (Continued from the previous page.)

(F) Overlaps of retrieved patterns over a wide range of network parameters. MF examples and PP concepts are retrieved at high and low threshold, respectively, optimized by grid search. Cue inaccuracy is the fraction of randomly chosen neurons in the target pattern whose activity is flipped to form the cue. Cue incompleteness is the fraction of randomly chosen active neurons in the target pattern which are inactivated to form the cue. Dendritic nonlinearity η introduces nonlinear summation between MF patterns $x_{\mu\nu}^{\text{MF}}$ and PP patterns $x_{\mu\nu}^{\text{PP}}$ by adding a term $\eta x_{\mu\nu i}^{\text{MF}} x_{\mu\nu i}^{\text{PP}}$ to Eq. 8 of the main text. Negative and positive η correspond to sublinear and superlinear regimes, respectively. For PP pattern strength $\gamma = 0.04, 0.08, 0.12, 0.16$, and 0.20 , we respectively use example loads $s = 400, 120, 80, 50$, and 20 ; otherwise, $s = 100$. Points represent means over 4 networks with 15 cues tested in each. (G, H) Similar to Fig. 3H, I, but for different MF pattern densities a_{MF} . PP patterns have correlation $\rho_{\text{PP}} = 0.04$. Source data are provided as a Source Data file.



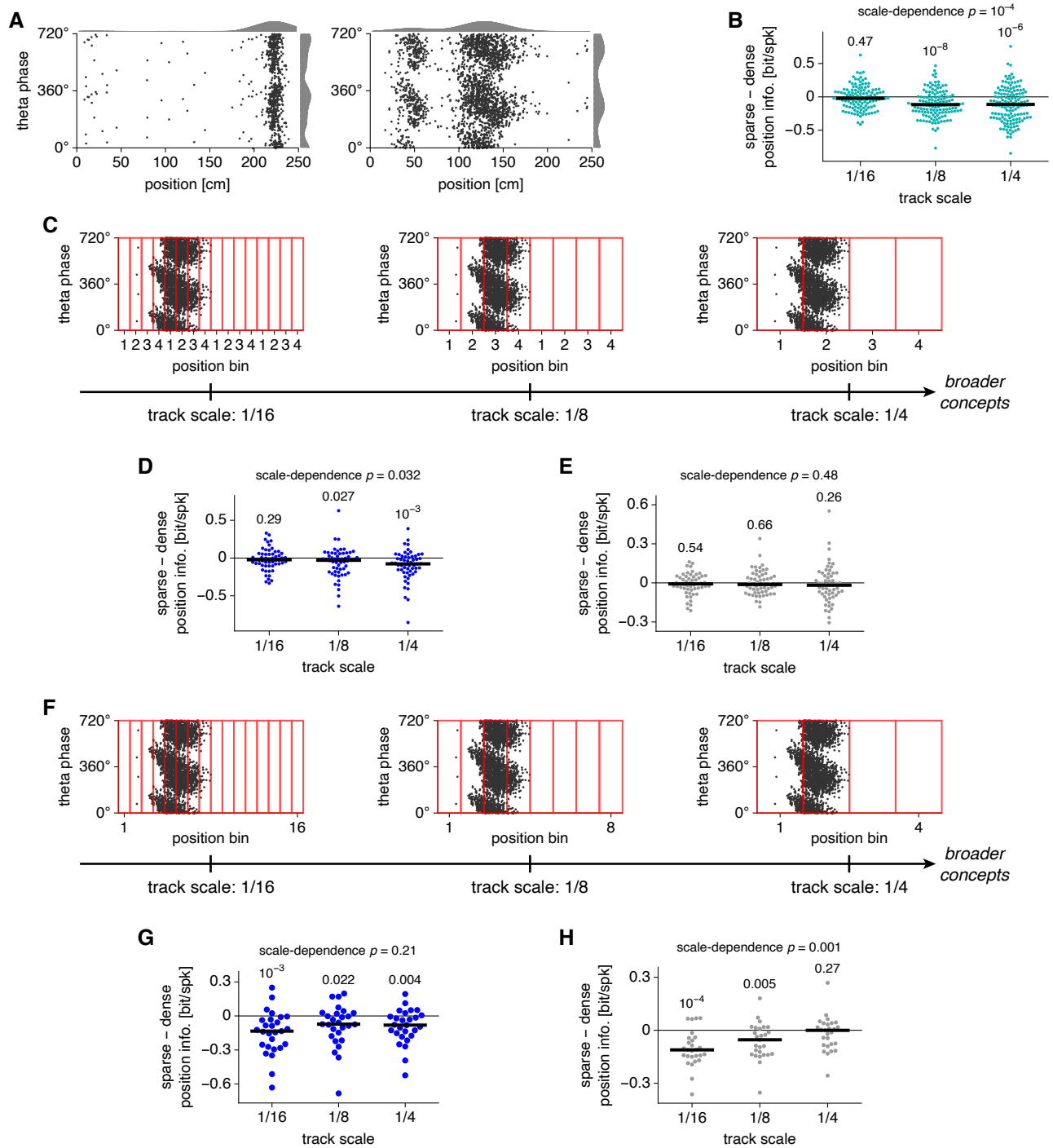
Supplementary Figure 4: Extended results for Fig. 4 of the main text. We use random MF and PP patterns instead of FashionMNIST encodings. (A) Similar to Fig. 4A. (B) Similar to Fig. 4C. For each scenario in B, 10 cues are tested in each of 10 networks. In all networks, 3 concepts are used, MF patterns have correlation 0, and PP patterns have density 0.5 and correlation 0.16. For MF densities $a_{\text{MF}} = 0.002, 0.01$, and 0.05 , we store 5, 10, and 20 patterns per concept, respectively. Source data are provided as a Source Data file.



Supplementary Figure 5: Extended results for Fig. 5 of the main text. (**A–C**) Additional single-neuron results. (**A**) Similar to Fig. 5F, but for three additional precessing fields from CA3. (**B**) Similar to Fig. 5G, but for the fields in **A**. (Continued on the next page.)

Supplementary Figure 5: *(Continued from the previous page.)*

(C) Similar to Fig. 5L, but comparing place fields in CA3 with those in CA1. CA3 $n = 35$ and CA1 $n = 47$. Numbers indicate p -values calculated by two-tailed Wilcoxon signed-rank tests for each population and by the two-tailed Mann-Whitney U test for the comparison between them. (D–J) Results for aggregate fields formed by accumulating spikes across phase precessing place fields. (D) Spikes aggregated across 57 CA3 place fields. (E) Activity (black), raw position information per spike (blue), and mean null-matched position information (gray) by theta phase for the aggregate field in D. (F, G) Similar to D, E, but for 55 CA1 place fields. (H) Similar to C, but comparing 100 bootstrap subsamples per region of 1000 spikes from the aggregate fields in D and F. (I) Parametric plots of features of the aggregate CA3 field in D with respect to theta phase. Field width and field center are respectively the standard deviation and mean of progress values over spikes. Each point represents one theta phase and its size is proportional to total activity. Arrows start at the phase with lowest activity and point towards increasing phase. (J) Similar to I, but for the aggregate CA1 field in F. For all results, spikes during each traveling direction are separately analyzed. In C, H, I, and J, information is sparsity-corrected with horizontal lines indicating medians. Source data are provided as a Source Data file.



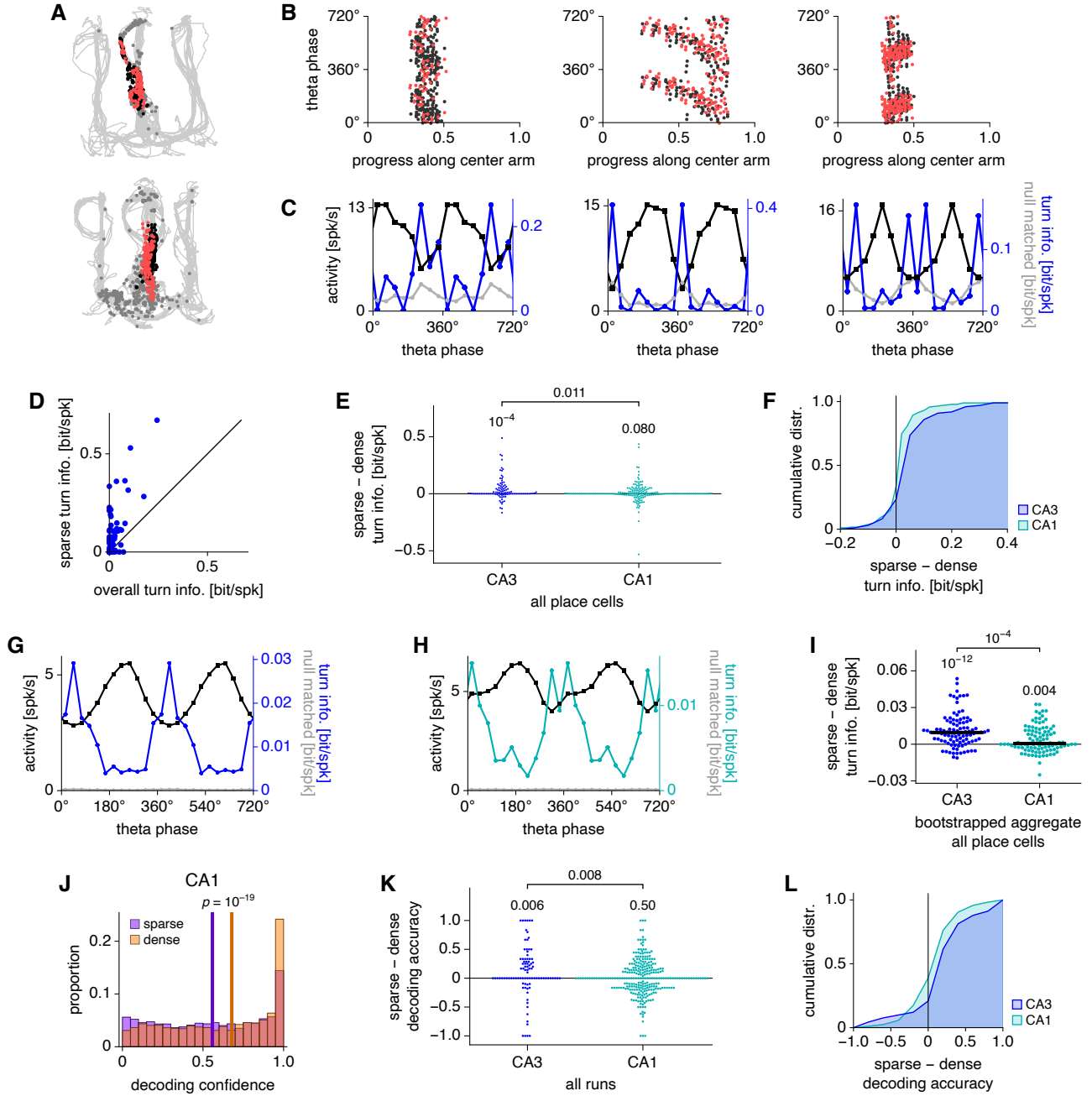
Supplementary Figure 6: Extended results for Fig. 6 of the main text. **(A)** Two additional CA3 place cells along a linear track. For each neuron, spikes are represented by two points at equivalent phases and are accumulated over position (top) and phase (right). **(B)** Similar to Fig. 6E, but for CA1 place cells. Track scale 1/16 $n = 122$, 1/8 $n = 137$, and 1/4 $n = 144$. **(C–E)** Similar to Fig. 6C, E, F, but for an alternative method for binning positions across track scales. **(C)** Four bins are still used for all scales, but bins cycle across the whole track. **(D)** For coarser scales, dense phases convey more position information per spike, as in Fig. 6E. Each track scale $n = 56$. **(E)** Shuffled data exhibit no relationship between position information and theta phase across track scales, as in Fig. 6F. **(F–H)** Similar to Fig. 6C, E, F, but for a third method for binning positions across track scales. **(F)** Different numbers of bins are used across scales.

(Continued on the next page.)

Supplementary Figure 6: *(Continued from the previous page.)*

(G) At all scales, dense phases convey more position information per spike, which differs from Fig. 6E. Each track scale $n = 27$.

(H) Shuffled data exhibit a relationship between position information and theta phase for finer scales, which differs from Fig. 6F and invalidates this binning method. For all results, spikes during each traveling direction are separately analyzed. In **B**, **D**, **E**, **G**, and **H**, information is sparsity-corrected with horizontal lines indicating medians. Source data are provided as a Source Data file.

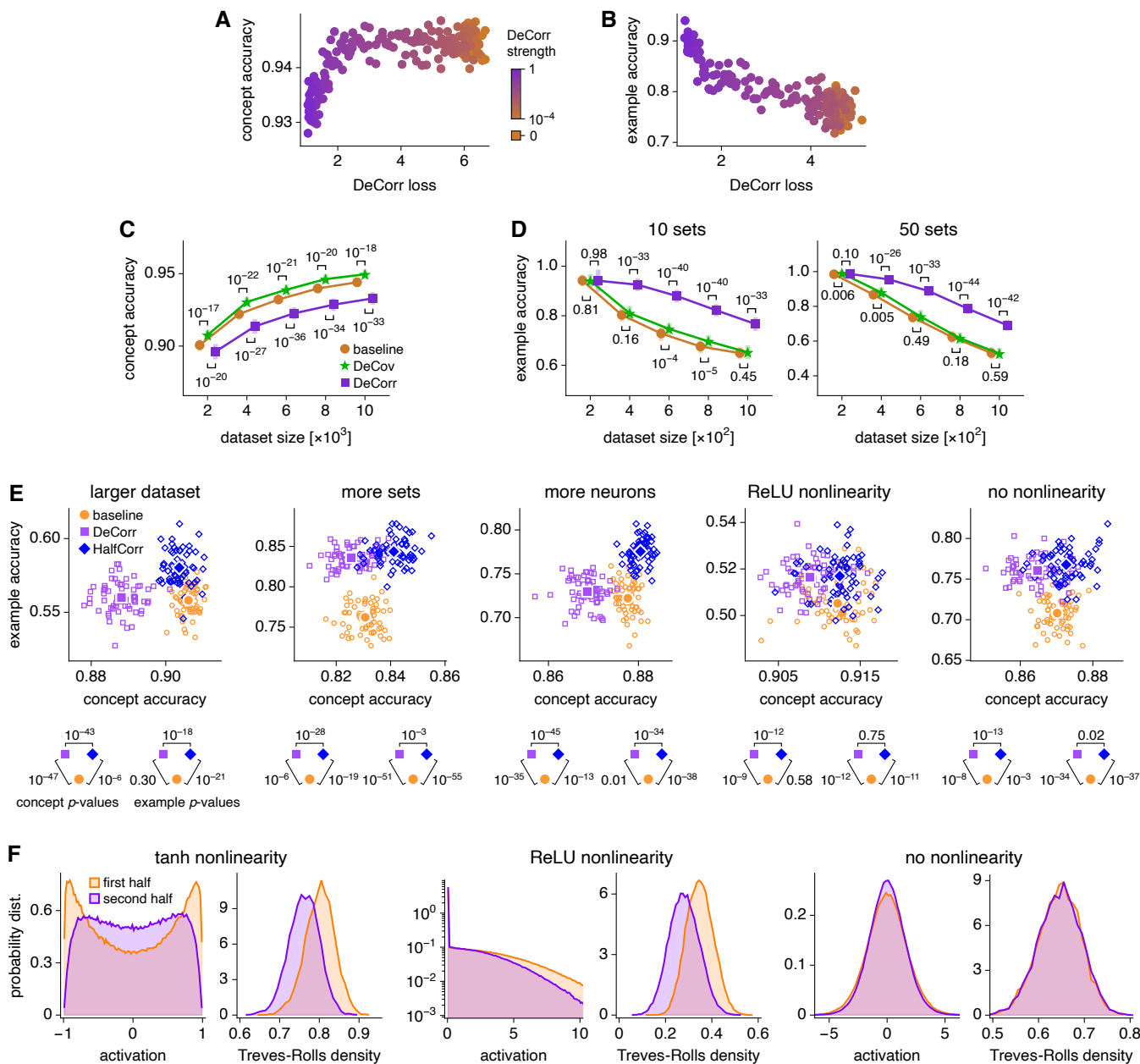


Supplementary Figure 7: Extended results for Fig. 7 of the main text. **(A–I)** Additional single-neuron results. **(A)** Spikes from Fig. 7C (top) and Fig. 7E (bottom) superimposed on the animal trajectory (light gray line) and other spikes (dark gray points). **(B)** Similar to Fig. 7C, E, but for three additional CA3 place cells. **(C)** Similar to Fig. 7D, F, but for the fields in **B**. **(D)** Average turn information per spike conveyed by CA3 place cells over sparse theta phases and over all phases. Each point represents one neuron, $n = 99$. Note that many neurons convey close to zero overall turn information but convey substantial sparse turn information. **(E)** Similar to Fig. 7G, but comparing place cells in CA3 with those in CA1. CA3 $n = 99$ and CA1 $n = 187$. Numbers indicate p -values calculated by two-tailed Wilcoxon signed-rank tests for each population by the two-tailed Mann-Whitney U test for the comparison between them. **(F)** Cumulative distribution functions for values in **E**. **(G)** Activity (black), raw position information per spike (blue), and mean null-matched position information (gray) by theta phase for spikes aggregated across 98 CA3 place cells. For each place cell, the turn direction with higher activity across all phases is identified. Aggregation is performed by collecting spikes corresponding to more active turn directions and those corresponding to less active directions. **(H)** Similar to **G**, but for 187 CA1 place cells.

(Continued on the next page.)

Supplementary Figure 7: *(Continued from the previous page.)*

(I) Similar to **E**, but comparing 100 bootstrap subsamples per region of 1000 spikes from the aggregates analyzed in **G** and **H**. (**J–L**) Additional Bayesian population decoding results. (**J**) Similar to Fig. 7J, but for CA1 place cells. (**K**) Similar to Fig. 7K, but comparing runs encoded by CA3 with those encoded by CA1. CA3 $n = 91$ and CA1 $n = 282$. (**L**) Cumulative distribution functions for values in **K**. For all results, spikes during each traveling direction are separately analyzed. In **D**, **E**, **F**, and **I**, information is sparsity-corrected with horizontal lines indicating medians. Source data are provided as a Source Data file.



Supplementary Figure 8: Extended results for Fig. 8 of the main text. **(A–D)** Additional results for the single-task architecture in Fig. 8B. **(A, B)** Concept and example accuracies as functions of DeCorr loss for 192 networks trained with various strengths of the DeCorr loss function. Increasing DeCorr strength decreases the final DeCorr loss, decreases concept accuracy, and increases example accuracy. Dataset sizes are respectively 10 000 and 500 in **A** and **B**, and 100 sets are used in **B**. **(C, D)** Similar to Fig. 8E, F, but including networks trained with the DeCov loss function developed by Michael Cogswell and colleagues¹. Unlike DeCorr, DeCov improves concept accuracy and does not substantially improve example accuracy compared to baseline. **(E, F)** Additional results for the multitask architecture in Fig. 8G. **(E)** Similar to Fig. 8I, but for different conditions. In each condition, HalfCorr networks exhibit the best combined performance. From left to right: dataset size of 3000 instead of 1000; 50 sets instead of 10; 500 neurons in each hidden layer instead of 100; ReLU activation function in each hidden layer instead of tanh; and linear activation in the second hidden layer and ReLU activation function in the first hidden layer, which makes the network equivalent to a single-layer perception. **(F)** Activation properties within the final hidden layer of HalfCorr networks with various activation functions described in **E**. Except for the linear activation case, the second, decorrelated half of the layer is sparser than the first, correlated half. Values elicited by 1000 train images in each of 8 trained networks. Treves-Rolls density is the *sparsity* defined in Treves and Rolls² and is computed with the absolute value of activations as in Willmore and Tolhurst³. Source data are provided as a Source Data file.

Supplementary Methods

Decorrelation in binary feedforward networks

Network architecture

We explore how the correlation of binary activity patterns changes when activity is propagated from one network to another. The two networks are termed presynaptic and postsynaptic. They have sizes N_{pre} and N_{post} . The presynaptic network exhibits activity patterns $x_{\nu i}^{\text{pre}} \in \{0, 1\}$, where $\nu = 1, \dots, s$ indexes patterns and $i = 1, \dots, N_{\text{pre}}$ indexes neurons.

W is the connectivity matrix from the presynaptic network to the postsynaptic network. For simplicity, we consider binary synaptic weights, so $W_{ij} \in \{0, 1\}$. The postsynaptic patterns $x_{\nu i}^{\text{post}} \in \{0, 1\}$ are determined by a simple threshold operation:

$$x_{\nu i}^{\text{post}} = \Theta \left[\sum_j W_{ij} x_{\nu j}^{\text{pre}} - \theta \right], \quad (1)$$

where Θ is the Heaviside step function and θ is the activity threshold. We can rewrite [Eq. 1](#) as

$$x_{\nu i}^{\text{post}} = \Theta[g_{\nu i} - \theta], \quad \text{where} \quad g_{\nu i} = \sum_j W_{ij} x_{\nu j}^{\text{pre}}. \quad (2)$$

Here, $g_{\nu i}$ is the synaptic input onto postsynaptic neuron i for pattern ν .

We characterize these patterns, either presynaptic or postsynaptic, by their density

$$a = \langle x_{\nu i} \rangle \quad (3)$$

and the correlation per neuron between two different patterns $\nu \neq \omega$

$$\rho = \frac{\langle x_{\nu i} x_{\omega i} \rangle - \langle x_{\nu i} \rangle \langle x_{\omega i} \rangle}{\langle x_{\nu i} x_{\nu i} \rangle - \langle x_{\nu i} \rangle \langle x_{\nu i} \rangle} = \frac{\langle x_{\nu i} x_{\omega i} \rangle - a^2}{a(1 - a)}. \quad (4)$$

The angle brackets indicate averages over patterns and neurons.

We now assume that the activity patterns and the connectivity matrix are generated via random processes. To be explicit, we will write $X_{\nu i}^{\text{pre}}$ as the random variable for the activity of presynaptic neuron i in pattern ν ; the same capitalization applies to the postsynaptic activities $X_{\nu i}^{\text{post}}$ and inputs $G_{\nu i}$. Lowercase letters represent samples of these variables.

We assume that each $X_{\nu i}^{\text{pre}}$ is an identically distributed random variable. We assume that the W_{ij} 's are independent and identically distributed (iid) Bernoulli random variables with parameter u :

$$W_{ij} \sim \text{Ber}(u). \quad (5)$$

In this case, each $X_{\nu i}^{\text{post}}$ is also an identically distributed random variable.

We can then write expressions for density and correlation as

$$\begin{aligned} a &= \mathbb{E}[X_{\nu i}] \\ \rho &= \frac{\mathbb{E}[X_{\nu i} X_{\omega i}] - a^2}{a(1 - a)}. \end{aligned} \quad (6)$$

These are population values; [Eqs. 3](#) and [4](#) indicate the sample estimates and should be written as $\hat{a}$ and $\hat{\rho}$,

but we will ignore this distinction.

Generating presynaptic activity patterns

For mathematical tractability, we will enforce density on a per-pattern basis; that is, each presynaptic pattern has the same number of active neurons:

$$n \equiv N_{\text{pre}} a_{\text{pre}} = \sum_i X_{\nu i}^{\text{pre}}. \quad (7)$$

We generate correlated patterns obeying this restriction as follows:

1. We create a concept pattern $\mathbf{x}^{\text{pre}} \in \{0, 1\}^{N_{\text{pre}}}$ by randomly choosing n neurons to be 1 and the rest to be 0. Let $\mathcal{S}$ be the set of all active neurons (value 1) and its complement $\mathcal{S}^c$ be the set of all inactive neurons (value 0).
2. To create each example pattern $\mathbf{x}_{\nu}^{\text{pre}}$, we randomly select a fraction d of neurons in $\mathcal{S}$ and set them to 0. We then randomly select the same number nd of neurons in $\mathcal{S}^c$ and set them to 1.

Now we calculate the correlation ρ_{pre} between the patterns $\mathbf{X}_{\nu}^{\text{pre}}$ generated by this method. To do so, we investigate the distribution of M , a random variable for the number of active neurons common to two different patterns. We note that

$$M = M_{\mathcal{S}} + M_{\mathcal{S}^c}, \quad (8)$$

where $M_{\mathcal{S}}$ is the number of neurons in $\mathcal{S}$ that remain active in both patterns, and $M_{\mathcal{S}^c}$ is the number of neurons in $\mathcal{S}^c$ that are activated in both patterns. These numbers have hypergeometric distributions:

$$\begin{aligned} M_{\mathcal{S}} &\sim \text{Hyp}[n, n(1-d), n(1-d)], \\ M_{\mathcal{S}^c} &\sim \text{Hyp}[N_{\text{pre}} - n, nd, nd]. \end{aligned} \quad (9)$$

We can then calculate

$$\mathbb{E}[X_{\nu i}^{\text{pre}} X_{\omega i}^{\text{pre}}] = \frac{\mathbb{E}[M]}{N_{\text{pre}}} = \frac{1}{N_{\text{pre}}} \left(\frac{n^2(1-d)^2}{n} + \frac{n^2 d^2}{N_{\text{pre}} - n} \right) = a_{\text{pre}}(1-d)^2 + \frac{a_{\text{pre}}^2 d^2}{1 - a_{\text{pre}}}. \quad (10)$$

Substituting this expression into [Eq. 6](#), we obtain

$$\begin{aligned} \rho_{\text{pre}} &= \left(\frac{1 - a_{\text{pre}} - d}{1 - a_{\text{pre}}} \right)^2, \\ d &= (1 - a_{\text{pre}})(1 - \sqrt{\rho_{\text{pre}}}). \end{aligned} \quad (11)$$

Thus, we know the number of activity flips required to produce presynaptic patterns with a given correlation.

Identifying the probability distribution of $G_{\nu i}$

Postsynaptic activity patterns are produced by [Eq. 2](#). Written in terms of random variables, it becomes

$$X_{\nu i}^{\text{post}} = \Theta[G_{\nu i} - \theta], \quad \text{where} \quad G_{\nu i} = \sum_j W_{ij} X_{\nu j}^{\text{pre}}. \quad (12)$$

Thus, the statistics of $\mathbf{X}_\nu^{\text{post}}$ are determined by $\mathbf{G}_\nu$. To calculate second-order statistics such as ρ_{post} , we need to determine the joint distribution of $G_{\nu i}$ and $G_{\omega i}$ for two patterns $\nu \neq \omega$.

To do so, we define $\mathcal{R}_0$ as the set of active neurons that are common to both patterns $\mathbf{X}_\nu^{\text{pre}}$ and $\mathbf{X}_\omega^{\text{pre}}$. We define $\mathcal{R}_\nu$ and $\mathcal{R}_\omega$ as sets of neurons only active in patterns ν and ω , respectively. We can then write

$$\begin{aligned} G_{\nu i} &= \sum_{j \in \mathcal{R}_0} W_{ij} + \sum_{j \in \mathcal{R}_\nu} W_{ij} \equiv G_{0i} + \tilde{G}_{\nu i} \\ G_{\omega i} &= \sum_{j \in \mathcal{R}_0} W_{ij} + \sum_{j \in \mathcal{R}_\omega} W_{ij} \equiv G_{0i} + \tilde{G}_{\omega i}. \end{aligned} \quad (13)$$

These sets have cardinalities

$$|\mathcal{R}_0| = M, \quad |\mathcal{R}_\nu| = n - M, \quad |\mathcal{R}_\omega| = n - M, \quad (14)$$

where n is given by Eq. 7 and M is given by Eqs. 8 and 9. Since the elements of $\mathbf{W}$ are iid Bernoulli random variables (Eq. 5),

$$\begin{aligned} (G_{0i} \mid M = m) &\sim \text{Bin}(m, u), \\ (\tilde{G}_{\nu i} \mid M = m) &\sim \text{Bin}(n - m, u), \\ (\tilde{G}_{\omega i} \mid M = m) &\sim \text{Bin}(n - m, u). \end{aligned} \quad (15)$$

Since $\mathcal{R}_0$, $\mathcal{R}_\nu$, and $\mathcal{R}_\omega$ are mutually disjoint, G_{0i} , $\tilde{G}_{\nu i}$, and $\tilde{G}_{\omega i}$ are mutually independent when conditioned on M .

Calculating the joint probability density function of $G_{\nu i}$ and $G_{\omega i}$

We take the $N_{\text{pre}} \rightarrow \infty$ limit in which the discrete probability distributions for $X_{\nu i}$ and $G_{\nu i}$ are replaced by their continuous limits. First, we present a few points on notation.

1. A continuous random variable A has probability density $f_A(a)$ at value $A = a$. We may simplify it as $f(a)$.
2. Variables A and B have joint probability density $f_{A,B}(a,b)$ at values $A = a$ and $B = b$. We may simplify it as $f(a,b)$.
3. Similarly, the conditional probability density given that the random variable B takes value b is $f_{A|B=b}(a)$. We may simplify it as $f(a|b)$.

We can now write the joint probability density function (pdf) for $G_{\nu i}$ and $G_{\omega i}$. For notational convenience, we will drop the subscript i for all relevant variables. The pdf is

$$\begin{aligned} f(g_\nu, g_\omega) &= \int dg_0 f(g_\nu, g_\omega, g_0) \\ &= \int dg_0 f_{\tilde{G}_\nu, \tilde{G}_\omega, G_0}(g_\nu - g_0, g_\omega - g_0, g_0) \\ &= \int dm \int dg_0 f_{\tilde{G}_\nu, \tilde{G}_\omega, G_0}(g_\nu - g_0, g_\omega - g_0, g_0 \mid m) f(m) \\ &= \int dm \int dg_0 f_{\tilde{G}_\nu}(g_\nu - g_0 \mid m) f_{\tilde{G}_\omega}(g_\omega - g_0 \mid m) f(g_0 \mid m) f(m). \end{aligned} \quad (16)$$

The second line is obtained using the change-of-variables formula for a joint pdf.

We have expressions for each pdf in the integrand. In the large N_{pre} limit, the binomial distributions in Eq. 15 can be approximated by normal distributions (dropping the subscript i for notational convenience):

$$\begin{aligned}(G_0 | M = m) &\sim \mathcal{N}[mu, mu(1-u)], \\ (\tilde{G}_\nu | M = m) &\sim \mathcal{N}[(n-m)u, (n-m)u(1-u)], \\ (\tilde{G}_\omega | M = m) &\sim \mathcal{N}[(n-m)u, (n-m)u(1-u)].\end{aligned}\tag{17}$$

Thus, we find

$$\begin{aligned}&\int dg_0 f_{\tilde{G}_\nu}(g_\nu - g_0|m) f_{\tilde{G}_\omega}(g_\omega - g_0|m) f(g_0|m) \\ &\propto \int dg_0 \exp\left[-\frac{(g_\nu - g_0 - (n-m)u)^2}{2(n-m)u(1-u)}\right] \exp\left[-\frac{(g_\omega - g_0 - (n-m)u)^2}{2(n-m)u(1-u)}\right] \exp\left[-\frac{(g_0 - mu)^2}{2mu(1-u)}\right] \\ &= \int dg_0 \exp\left[-\frac{[g_0 - (\frac{g_\nu + g_\omega}{2} - (n-m)u)]^2}{(n-m)u(1-u)}\right] \exp\left[-\frac{(\frac{g_\nu - g_\omega}{2})^2}{(n-m)u(1-u)}\right] \exp\left[-\frac{(g_0 - mu)^2}{2mu(1-u)}\right] \\ &\propto \exp\left[-\frac{(\frac{g_\nu + g_\omega}{2} - nu)^2}{(n+m)u(1-u)}\right] \exp\left[-\frac{(\frac{g_\nu - g_\omega}{2})^2}{(n-m)u(1-u)}\right].\end{aligned}\tag{18}$$

We can write the terms inside the exponential as

$$\begin{aligned}&\frac{(\frac{g_\nu + g_\omega}{2} - nu)^2}{(n+m)u(1-u)} + \frac{(\frac{g_\nu - g_\omega}{2})^2}{(n-m)u(1-u)} \\ &= \frac{\left[\frac{(g_\nu - nu) + (g_\omega - nu)}{2}\right]^2}{(n+m)u(1-u)} + \frac{\left[\frac{(g_\nu - nu) - (g_\omega - nu)}{2}\right]^2}{(n-m)u(1-u)} \\ &= \frac{1}{2} \left[\begin{pmatrix} g_\nu - nu & g_\omega - nu \end{pmatrix} \frac{1}{2u(1-u)} \begin{pmatrix} \frac{1}{n+m} + \frac{1}{n-m} & \frac{1}{n+m} - \frac{1}{n-m} \\ \frac{1}{n+m} - \frac{1}{n-m} & \frac{1}{n+m} + \frac{1}{n-m} \end{pmatrix} \begin{pmatrix} g_\nu - nu \\ g_\omega - nu \end{pmatrix} \right] \\ &= \frac{1}{2} (\mathbf{g} - \boldsymbol{\mu}_G)^\top \boldsymbol{\Sigma}_G^{-1} (\mathbf{g} - \boldsymbol{\mu}_G).\end{aligned}\tag{19}$$

The last expression is written in terms of the variable vector, mean vector, and covariance matrix for G_ν and G_ω :

$$\begin{aligned}\mathbf{g} &= \begin{pmatrix} g_\nu \\ g_\omega \end{pmatrix} \\ \boldsymbol{\mu}_G &= \begin{pmatrix} nu \\ nu \end{pmatrix} = \begin{pmatrix} N_{\text{pre}} a_{\text{pre}} u \\ N_{\text{pre}} a_{\text{pre}} u \end{pmatrix} \\ \boldsymbol{\Sigma}_G &= \begin{pmatrix} nu(1-u) & mu(1-u) \\ mu(1-u) & nu(1-u) \end{pmatrix} = \sigma_G^2 \begin{pmatrix} 1 & \rho_G \\ \rho_G & 1 \end{pmatrix},\end{aligned}\tag{20}$$

where the covariance and correlation are

$$\sigma_G^2 = N_{\text{pre}} a_{\text{pre}} u(1-u) \quad \text{and} \quad \rho_G = \frac{m}{N_{\text{pre}} a_{\text{pre}}}.\tag{21}$$

Combining Eqs. 16, 18, and 19, we obtain

$$f(g_\nu, g_\omega) \propto \int dm \exp \left[-\frac{1}{2} (\mathbf{g} - \boldsymbol{\mu}_G)^\top \boldsymbol{\Sigma}_G^{-1} (\mathbf{g} - \boldsymbol{\mu}_G) \right] f(m). \quad (22)$$

Now we consider $M = M_S + M_{S^c}$ (Eq. 8). From Eq. 9, we see that M_S and M_{S^c} have means and variances

$$\begin{aligned} \mu_S &= n(1-d)^2 &= N_{\text{pre}} a_{\text{pre}} (1-d)^2 \\ \sigma_S^2 &= nd^2(1-d)^2 &= N_{\text{pre}} a_{\text{pre}} d^2 (1-d)^2 \\ \mu_{S^c} &= \frac{n^2 d^2}{N_{\text{pre}} - n} &= N_{\text{pre}} \frac{a_{\text{pre}}^2 d^2}{1 - a_{\text{pre}}} \\ \sigma_{S^c}^2 &= \frac{n^2 d^2 (N_{\text{pre}} - n(1+d))^2}{(N_{\text{pre}} - n)^3} &= N_{\text{pre}} \frac{a_{\text{pre}}^2 d^2 (1 - a_{\text{pre}} - a_{\text{pre}} d)^2}{(1 - a_{\text{pre}})^3}. \end{aligned} \quad (23)$$

Note that the flip fraction d can be expressed in terms of a_{pre} and ρ_{pre} via Eq. 11.

As $N_{\text{pre}} \rightarrow \infty$, the hypergeometric random variables M_S and M_{S^c} approach normal distributions with the means and variances in Eq. 23. Thus, their distributions become sharply peaked around their means, and we can approximate the pdf of M by a delta-function at its mean:

$$\begin{aligned} f(m) &\rightarrow \delta(m - \mu_M), \quad \text{where} \\ \mu_M &= N_{\text{pre}} a_{\text{pre}} (1-d)^2 + N_{\text{pre}} \frac{a_{\text{pre}}^2 d^2}{1 - a_{\text{pre}}} \\ &= N_{\text{pre}} a_{\text{pre}} (a_{\text{pre}} + \rho_{\text{pre}} - a_{\text{pre}} \rho_{\text{pre}}). \end{aligned} \quad (24)$$

We now have our final expression for the joint pdf of $G_{\nu i}$ and $G_{\omega i}$. Reintroducing the neural index i and the normalization factor, Eq. 22 becomes

$$f(g_{\nu i}, g_{\omega i}) = \frac{1}{2\pi \sqrt{\det \boldsymbol{\Sigma}_G}} \exp \left[-\frac{1}{2} (\mathbf{g}_i - \boldsymbol{\mu}_G)^\top \boldsymbol{\Sigma}_G^{-1} (\mathbf{g}_i - \boldsymbol{\mu}_G) \right], \quad (25)$$

where

$$\mathbf{g}_i = \begin{pmatrix} g_{\nu i} \\ g_{\omega i} \end{pmatrix}, \quad \boldsymbol{\mu}_G = \begin{pmatrix} \mu_G \\ \mu_G \end{pmatrix}, \quad \boldsymbol{\Sigma}_G = \sigma_G^2 \begin{pmatrix} 1 & \rho_G \\ \rho_G & 1 \end{pmatrix} \quad (26)$$

and

$$\mu_G = N_{\text{pre}} a_{\text{pre}} u, \quad \sigma_G^2 = N_{\text{pre}} a_{\text{pre}} u(1-u), \quad \rho_G = a_{\text{pre}} + \rho_{\text{pre}} - a_{\text{pre}} \rho_{\text{pre}}. \quad (27)$$

Supplementary Figure 1 shows a plot of the joint pdf $f(g_{\nu i}, g_{\omega i})$ along with a histogram obtained through numerical simulation. The theoretical formula Eq. 25 agrees very well with the numerical data.

Integrating the joint probability density function to obtain a_{post} and ρ_{post}

With the joint pdf for $G_{\nu i}$ and $G_{\omega i}$ (Eq. 25), we can compute the postsynaptic pattern density a_{post} and correlation ρ_{post} using Eqs. 6 and 12. According to Eq. 12, $X_{\nu i}^{\text{post}}$ acts as an indicator random variable for $G_{\nu i} > \theta$, and the product $X_{\nu i}^{\text{post}} X_{\omega i}^{\text{post}}$ acts as an indicator random variable for $G_{\nu i} > \theta \cap G_{\omega i} > \theta$. Thus,

$$\mathbb{E}[X_{\nu i}^{\text{post}}] = P(G_{\nu i} > \theta) \quad \text{and} \quad \mathbb{E}[X_{\nu i}^{\text{post}} X_{\omega i}^{\text{post}}] = P(G_{\nu i} > \theta \cap G_{\omega i} > \theta). \quad (28)$$

We can calculate

$$\begin{aligned}
\mathbb{E}[X_{\nu i}^{\text{post}}] &= \int_{\theta}^{\infty} dg_{\nu i} \int_{-\infty}^{\infty} dg_{\omega i} f(g_{\nu i}, g_{\omega i}) \\
&= \int_{\theta}^{\infty} dg_{\nu i} \frac{1}{\sqrt{2\pi}\sigma_G} \exp\left[-\frac{(g_{\nu i} - \mu_G)^2}{2\sigma_G^2}\right] \\
&= \frac{1}{2} \operatorname{erfc} \frac{\theta - \mu_G}{\sqrt{2}\sigma_G}.
\end{aligned} \tag{29}$$

Thus, the postsynaptic pattern density is immediately

$$a_{\text{post}} = \mathbb{E}[X_{\nu i}^{\text{post}}] = \frac{1}{2} \operatorname{erfc} \frac{\phi}{\sqrt{2}}, \quad \text{where } \phi = \frac{\theta - \mu_G}{\sigma_G}. \tag{30}$$

The rescaled threshold ϕ is the standardized version of θ .

We next need to calculate

$$\begin{aligned}
\mathbb{E}[X_{\nu i}^{\text{post}} X_{\omega i}^{\text{post}}] &= \int_{\theta}^{\infty} dg_{\nu i} \int_{\theta}^{\infty} dg_{\omega i} f(g_{\nu i}, g_{\omega i}) \\
&= \frac{1}{2\pi\sqrt{\det \mathbf{\Sigma}_G}} \int_{\theta}^{\infty} dg_{\nu i} \int_{\theta}^{\infty} dg_{\omega i} \exp\left[-\frac{1}{2}(\mathbf{g}_i - \boldsymbol{\mu}_G)^{\top} \mathbf{\Sigma}_G^{-1} (\mathbf{g}_i - \boldsymbol{\mu}_G)\right].
\end{aligned} \tag{31}$$

By standardizing the variables of integration with $h_{\nu i} = (g_{\nu i} - \mu_G)/\sigma_G$, this integral can be expressed in terms of the standard bivariate normal:

$$\mathbb{E}[X_{\nu i}^{\text{post}} X_{\omega i}^{\text{post}}] = \frac{1}{2\pi\sqrt{1-\rho_G^2}} \int_{\phi}^{\infty} dh_{\nu i} \int_{\phi}^{\infty} dh_{\omega i} \exp\left[-\frac{h_{\nu i}^2 + h_{\omega i}^2 - 2\rho_G h_{\nu i} h_{\omega i}}{2(1-\rho_G^2)}\right]. \tag{32}$$

This double integral cannot be evaluated in closed form, but we can reduce it to a single integral⁴:

$$\mathbb{E}[X_{\nu i}^{\text{post}} X_{\omega i}^{\text{post}}] = \Gamma[\phi, \rho_G] \equiv \frac{1}{2\pi} \int_{\arccos \rho_G}^{\pi} d\psi \exp\left[-\frac{\phi^2}{1 + \cos \psi}\right]. \tag{33}$$

Therefore, the expression for the postsynaptic correlation follows:

$$\rho_{\text{post}} = \frac{\Gamma[\phi, \rho_G] - a_{\text{post}}^2}{a_{\text{post}}(1 - a_{\text{post}})}, \tag{34}$$

where a_{post} can be expressed in terms of the standardized threshold ϕ with [Eq. 30](#). On the other hand, we can stipulate a desired a_{post} and then recover ϕ and ρ_{post} with

$$\begin{aligned}
\phi &= \sqrt{2} \operatorname{erfc}^{-1}(2a_{\text{post}}), \\
\rho_{\text{post}} &= \frac{\Gamma[\sqrt{2} \operatorname{erfc}^{-1}(2a_{\text{post}}), a_{\text{pre}} + \rho_{\text{pre}} - a_{\text{pre}}\rho_{\text{pre}}] - a_{\text{post}}^2}{a_{\text{post}}(1 - a_{\text{post}})}.
\end{aligned} \tag{35}$$

This is Eq. 1 of the main text. Figure 2E shows that this formula for the postsynaptic correlation agrees well with values obtained through numerical simulation across a variety of parameter values.

Exploring ρ_{post} as a function of a_{pre} , a_{post} , and ρ_{pre}

In [Supplementary Fig. 2B](#), we plot ρ_{post} as a function a_{pre} and a_{post} for various ρ_{pre} . We see that decorrelation ($\rho_{\text{post}} < \rho_{\text{pre}}$) occurs when a_{post} is low. Thus, a downstream (postsynaptic) network with many neurons but low activity naturally decorrelates patterns of the upstream (presynaptic) network. The low activity can be achieved by low connectivity u or a high threshold θ ; [Eq. 35](#) does not differentiate between the two.

[Supplementary Figure 2C](#) demonstrates that if presynaptic patterns are sparse and decorrelated (lower left corner of the plot), postsynaptic patterns also exhibit low correlation even if they are denser. Thus, once patterns are sparsified and decorrelated, they will remain decorrelated for subsequent feedforward layers. Note that this panel also shows the symmetry in interchanging $a_{\text{pre}} \leftrightarrow \rho_{\text{pre}}$ present in [Eq. 35](#).

CA3 model with random binary patterns

We use a network size of $N_{\text{CA3}} = 10\,000$. We generate MF example patterns $\mathbf{x}_{\mu\nu}^{\text{MF}}$ with desired density a_{MF} and correlation 0 by randomly activating $N_{\text{CA3}}a_{\text{MF}}$ neurons. We generate PP concept patterns $\mathbf{x}_{\mu}^{\text{PP}}$ with desired density 0.5 by randomly activating each neuron with probability 0.5. We then generate PP example patterns $\mathbf{x}_{\mu\nu}^{\text{PP}}$ with desired correlation ρ_{PP} by randomly flipping each concept neuron with probability $(1 - \sqrt{\rho_{\text{PP}}})/2$. Simulations are initiated without cue noise and a sharp activation threshold ($\beta \rightarrow \infty$) to assess the best possible network performance.

To calculate capacities in [Fig. 3H, I](#), we use a higher strength of PP inputs $\zeta = 0.2$ to make the capacity values more computationally accessible. For a given load of concepts per neuron, we perform a grid search over the load of examples per concept using 8 networks per load and testing 20 cues in each network. Each cue is identical to its target pattern. We search for the load at which the average overlap crosses a threshold, which is $1/2$, $(1 + \rho_{\text{PP}})/2$, and $(1 + \sqrt{\rho_{\text{PP}}})/2$ for MF examples, PP examples, and PP concepts, respectively, to account for the positive overlap of off-target PP patterns if they are correlated⁵. For MF examples, we explore activity thresholds θ' between 0.43 and 0.85 and use the value that maximizes average overlap. For PP examples and concepts, we set the activity threshold θ' to 0.

CA3 model behavior during oscillating threshold

The oscillation analysis in [Fig. 4C](#) characterizes network behavior between update cycles 60 and 120. Consider a single oscillation cycle. For example-related behavior, we consider its high-threshold half. Let $m_1(t)$ and $m_2(t)$ be the largest and second-largest overlaps within the target concept at time t , and let $m_0(t)$ be the largest overlap within other concepts. If $m_1(t) > 0.8$, $m_1(t) > 2 \cdot m_2(t)$, and $m_1(t) > 2 \cdot m_0(t)$ at any time, the behavior of the oscillation cycle is categorized as *single examples within a target concept*. If $m_1(t) > 0.8$, $m_1(t) < 2 \cdot m_2(t)$, and $m_1(t) > 2 \cdot m_0(t)$, the behavior is categorized as *mixed examples within a target concept*. If at least half of the oscillation cycles receive a certain categorization, it is considered the network behavior. Otherwise, the network behavior is *examples within other concepts*. For concept-related behavior, we consider the low-threshold half of each oscillation cycle. Let $m_1(t)$ be the overlap with the target concept and $m_0(t)$ be the largest overlap with other concepts. If $m_1(t) > 0.1$ and $m_1(t) > 2 \cdot m_0(t)$ at any time, the behavior of the oscillation cycle is categorized as *target concept* (for the random binary patterns in [Supplementary Fig. 4B](#), we use $m_1(t) > 0.2$ instead). If at least half of the oscillation cycles receive this categorization, it is considered the network behavior. Otherwise, the network behavior is *other concepts*.

Experimental data preprocessing

In all experimental analyses, we consider each traveling direction separately. Thus, each recorded neuron effectively yields two neurons in our analyses with their own spikes and trajectory occupancies.

Linear track data from the CRCNS hc-3 dataset is used to produce the results in Figs. 5 and 6⁶. For CA3, we use all linear track sessions from rats ec013, ec016, gor, and vvp with CA3 neurons recorded, and for CA1, we use all linear track sessions between 761 and 882 from rat ec013. For rats gor and vvp, the size of the linear track was changed during recording, and we only consider data before the change. Animal positions are taken to be the mean of the two LED lights on the microdrive. The track axis is taken to be the first principal component of sampled positions. Animal velocities are differences between tracked position samples divided by the sampling rate and smoothed with a Gaussian filter with standard deviation 0.1 s.

We use the recommended quality criteria involving eDist, RefracRatio, and RefracViol when selecting units. Since we are interested in the theta oscillation, we only consider spikes occurring during locomotion with speed greater than 10 cm/s. We require units to have at least 50 spikes occurring within the central 70% of the track to avoid neurons whose behavior may be dominated by boundary effects. To determine the theta signal for each unit, we identify the tetrode from which it was recorded and average the LFP over all channels on the tetrode. This signal is bandpass-filtered between 6–10 Hz, and the complex argument of its Hilbert transform is the local theta phase.

To extract place fields for Fig. 5, we first discretize track positions with 1 cm bins. For a given place cell, we first compute the activity across all theta phases as a function of position and apply a Gaussian filter whose standard deviation is 0.01 times the track length. We find activity peaks whose maximum is 0.6 standard deviations above the mean; if the activity exhibits multiple peaks while remaining above this threshold, the largest is chosen. We then find the closest flanking positions where the activity falls below 0.2 times the peak value. The region in between is the place field. If two place fields overlap, they are divided at the activity minimum located between their peaks.

W-maze data from the CRCNS hc-6 dataset is used to produce the results in Fig. 7⁷. We use all run sessions from rats bon, con, dud, fra, mil, and ten. We remove 2 sessions from dud and 3 sessions from fra without position samples in one of the side arms. Animal positions and velocities are taken directly as the 30 Hz-interpolated samples from the dataset. We only consider spikes occurring during locomotion with speed greater than 5 cm/s. We require units to have at least 30 spikes occurring within the center arm. Spike theta phases are taken directly from the dataset.

To extract arm identity from the position samples, we first linearly rescale both position coordinates to span from 0 to 1. We define a maze skeleton consisting of lines that represent the left arm from (0.1, 0.1) to (0.1, 1), the center arm from (0.5, 0.1) to (0.5, 1), the right arm from (0.9, 0.1) to (0.9, 1), and the base from (0.1, 0.1) to (0.9, 0.1). We then fit transformations of this skeleton to the position samples, allowing for stretching along the first coordinate, rotation about its center, and translations. The fit is performed by minimizing the total squared distance between each sampled position and its closest point on the transformed skeleton. Empirically, this process yields excellent fitting without any need for manual intervention. Then, each position sample is assigned to either the left, center, or right arm based on closest distance.

To extract runs along the central arm, we first smooth arm samples by encoding arm identity as a one-hot vector, applying a Gaussian filter with standard deviation 0.2 s, and identifying the largest element in each sample. We then consider each span of central-arm samples. We find the times within the span at which the animal crosses scaled positions 0.4 and 0.7, where 0 corresponds to the smallest position value at the base of the maze and 1 corresponds to the largest position value at the far end of the arms. The run duration

must be between 0.1 s and 10 s, and we then pad the run by 0.5 s at both ends. Outward runs cross scaled position 0.7 before 0.4, and the subsequent arm identity is used to determine the future turn direction. Inward runs cross scaled position 0.4 before 0.7, and the previous arm identity is used to determine the past turn direction.

Experimental data aggregate analysis

The aggregate fields in [Supplementary Fig. 5D–J](#) are formed from phase-precessing place fields. We do not enforce a minimum spike count or ensure theta modulation on a single-neuron basis. In total, we collect 19 678 spikes from 57 CA3 place fields and 29 664 spikes from 55 CA1 place fields. We perform bootstrapping by sampling with replacement 1000 spikes at a time. For each subsample, we bin spikes into 10 progress bins and phase bins of width 15° .

The aggregate fields in [Supplementary Fig. 7G–I](#) are formed from place cells with at least 30 spikes within the central arm. We do not ensure theta modulation on a single-neuron basis. For each neuron, we identify the turn directions with higher and lower activities, and collect spikes occurring during each condition. In total, we collect 47 196 spikes from 331 CA3 place cells and 72 461 spikes from 436 CA1 place cells. We perform bootstrapping by sampling with replacement 1000 spikes at a time. For each subsample, we bin spikes into 2 directions (more active and less active) and phase bins of width 30° .

References

- [1] M. Cogswell, F. Ahmed, R. Girshick, L. Zitnick, and D. Batra. Reducing overfitting in deep networks by decorrelating representations. *arXiv* 1511.06068, 2015.
- [2] A. Treves and E. T. Rolls. What determines the capacity of autoassociative memories in the brain? *Netw. Comput. Neural Syst.*, 2(4):371–397, 1991.
- [3] B. Willmore and D. J. Tolhurst. Characterizing the sparseness of neural codes. *Netw. Comput. Neural Syst.*, 12(3):255–270, 2001.
- [4] D. B. Owen. Tables for computing bivariate normal probabilities. *Ann. Math. Stat.*, 27(4):1075–1090, 1956.
- [5] L. Kang and T. Toyozumi. Hopfield-like network with complementary encodings of memories. *Phys. Rev. E*, 108(5):054410, 2023.
- [6] K. Mizuseki, A. Sirota, E. Pastalkova, K. Diba, and G. Buzsáki. Multiple single unit recordings from different rat hippocampal and entorhinal regions while the animals were performing multiple behavioral tasks. *CRCNS.org*, 2013.
- [7] M. Karlsson, M. Carr, and L. M. Frank. Simultaneous extracellular recordings from hippocampal areas CA1 and CA3 (or MEC and CA1) from rats performing an alternation task in two W-shaped tracks that are geometrically identically but visually distinct. *CRCNS.org*, 2015.