Symmetries and Continuous Attractors in Disordered Neural Circuits

David G. Clark[1,2], L.F. Abbott[1,2], and Haim Sompolinsky[3,4]

dgc2138@cumc.columbia.edu, lfa2103@columbia.edu, hsompolinsky@mcb.harvard.edu

[1]*Zuckerman Institute, Columbia University, New York, NY 10027, USA*
[2]*Kavli Institute for Brain Science, Columbia University, New York, NY 10027, USA*
[3]*Center for Brain Science, Harvard University, Cambridge, MA 02138, USA*
[4]*Edmond and Lily Safra Center for Brain Sciences, Hebrew University of Jerusalem, Jerusalem 9190401, Israel*

Abstract

A major challenge in neuroscience is reconciling idealized theoretical models with complex, heterogeneous experimental data. We address this challenge through the lens of continuous-attractor networks, which model how neural circuits may represent continuous variables, such as head direction or spatial location, through collective dynamics. In classical continuous-attractor models, a continuous symmetry of the connectivity generates a manifold of stable states, resulting in tuning curves that are identical up to shifts. However, mouse head-direction cells show substantial heterogeneity in their responses that appears incompatible with this classical picture. To understand the mechanistic origin of these data, we use an optimization principle to construct recurrent-network models that match the observed responses while exhibiting quasi-continuous-attractor dynamics. To study how such systems scale with increasing numbers of neurons N, we develop a statistical generative process that produces artificial tuning curves that match many features of the experimental data quantitatively. Due to a continuous symmetry in the generative process, the connectivity matrix exhibits doublet degeneracy in its spectrum at large N, reflecting an underlying circular geometry. Unlike classical models, where the ring structure is embedded in the neuronal space through a structured Fourier embedding, our model uses a random, disordered embedding. Analysis of the network dynamics in the large-N limit through dynamical mean-field theory reveals that the system becomes equivalent to a classical ring-attractor model. We extend this approach to higher-dimensional symmetries, applying it to grid cells in the medial entorhinal cortex and showing that the mean-field description recovers classical continuous-attractor models. Our work implies that large mammalian neural circuits could represent continuous variables using continuous-attractor dynamics, complete with continuous symmetry, in a way that is fully compatible with their heterogeneity and disorder.

1 Introduction

Organisms from flies to humans maintain mental representations of continuous variables. These include navigationally relevant variables, such as self-orientation and location, as well as more abstract cognitive variables such as decision confidence. Continuous-attractor models explain how networks of interacting neurons, each individually lacking memory, can maintain these representations through their collective dynamics. In these models, synaptic connectivity is configured to generate a manifold of stable states that matches the dimension and topology of the encoded variable, providing persistent working memory of this variable. Furthermore, by incorporating velocity inputs that modify network states, such systems can integrate these inputs and compute changing values of the encoded continuous variable.

Classical continuous-attractor models generate marginally stable manifolds of fixed points through three essential ingredients (Fig. 1A, B; see [1]). First, neurons must transform input currents into firing rates in a nonlinear manner. Second, the connectivity must support Turing pattern formation, typically through a Mexican-hat profile involving local excitation and longer-range inhibition. Third, the connectivity must possess a continuous symmetry corresponding to the encoded variable—for example, one-dimensional circular symmetry for ring attractors representing head direction, or two-dimensional symmetry for toroidal attractors representing grid-cell responses. This symmetry ensures that when pattern formation spontaneously breaks the continuous symmetry, the resulting activity patterns form a continuous manifold with the same dimension and topology as the connectivity and thus as the encoded variable. As a consequence of this symmetry, the tuning curves of individual neurons (i.e., their responses as a function of the encoded variable) are identical up to neuron-specific shifts (Fig. 1C).

A canonical example of continuous-attractor models is the ring attractor for one-dimensional circular variables, such as head direction or visual-stimulus orientation [2–8]. The experimental discovery of a ring-attractor structure in the brain of *Drosophila melanogaster* several decades after the formulation of this model provided a notable validation of theoretical approaches. In the central complex, a region called the ellipsoid body has an anatomical ring structure that produces a localized bump of activity that accurately tracks the animal's orientation, appearing to be a nearly exact instantiation of a classical ring-attractor model [9–11]. Ring-attractor-like circuitry has also been discovered in zebrafish [12].

In this paper, we investigate how mammalian brains might implement stable representations of continuous variables. Rodents, for example, possess neurons that encode head direction (analogous to the heading representation in flies) in regions including the anterodorsal nucleus of the thalamus and the postsubiculum [13–15]. Similarly, recordings from the medial entorhinal cortex have revealed grid cells that represent spatial location through hexagonally arranged firing fields, with each cell's activity peaking at the vertices of a hexagonal lattice that spans the environment [16, 17]. An open question is whether these are examples of continuous-attractor mechanisms representing continuous variables. Although detailed synaptic connectivity in mammalian systems remains largely unknown, evidence suggests that the intrinsic circuit structure constrains neuronal activity to low-dimensional manifolds. Manifold confinement persists even during sleep—a hallmark prediction of continuous-attractor models. Such structured activity patterns have been demonstrated in both rodent head-direction cells, where activity is confined to a manifold with ring topology [18, 19], and grid cells, where activity traces out a manifold with toroidal topology [20–22].

While mammalian neural circuits possess the requisite nonlinear neurons and strong excitatory and inhibitory recurrent connectivity, they notably deviate from the orderly response properties characteristic of classical continuous-attractor models. Instead of having tuning curves that are identical up to neuron-specific shifts, mammalian circuits exhibit disordered responses, suggesting an absence of continuous symmetry in the connectivity. This heterogeneity is evident in the rodent head-direction system, which we will examine in detail. Similar disorder is seen in the medial entorhinal cortex, where neurons display a continuous spectrum of "grid scores" in their spatial firing patterns [23–25]. Indeed, only a small subset of grid cells exhibits the periodicity predicted by classical continuous-attractor models.

These observations pose a fundamental question: *How can theoretical models requiring explicit symmetry be reconciled with the disorder observed in biological circuits?* Despite the theoretical importance of this question, no theoretical framework has yet been proposed that conceptually reconciles continuous-attractor dynamics with the heterogeneous neuronal responses observed in experimental data. We investigate this question through analysis of the rodent head-direction system. By examining recordings from head-direction cells in the postsubiculum of freely moving mice [19], we first show that the heterogeneity in neuronal tuning curves violates the assumptions of classical ring-attractor models. To bridge this gap between theory and data, we use an optimization procedure to construct recurrent neural networks directly from experimentally observed tuning curves. These data-derived networks reproduce the heterogeneous tuning properties observed in the data while exhibiting quasi-continuous-attractor dynamics.

To extend this optimization procedure beyond the limited population sizes available in experimental recordings, we introduce a statistical generative process for constructing tuning curves. It is based on a Gaussian process that uses a circular symmetry suggested by the statistical distribution of experimentally observed tuning curves. This generative process reproduces the complex statistical structure of biological head-direction cells with remarkable accuracy across multiple quantitative measures.

Using the generative process, we can study the properties of these networks at scales reflecting mammalian neural circuits. Analysis of the optimized connectivity matrix reveals that, despite its apparent disorder, it possesses a circular symmetry that manifests in the form of paired (doubly degenerate) eigenvalues. However, this symmetry is realized in a different way than in classical ring-attractor models. While classical models embed their circular symmetry using orderly, Fourier patterns across neurons, our model achieves the same symmetry through random, disordered patterns, explaining how heterogeneous neural responses can coexist with ordered computation and informing how continuous attractors might be uncovered in connectome data.

Finally, to analyze the dynamics of large, synthetic data-derived networks, we develop a dynamical mean-field theory in the limit of infinitely many neurons. This reveals a central result: the network dynamics become mathematically equivalent to those of a classical ring-attractor model, exhibiting manifest continuous symmetry, even while individual neurons maintain their characteristic heterogeneous tuning. The mean-field description recovers all of the essential features of classical continuous attractors, including Mexican-hat connectivity supporting pattern formation and spontaneous symmetry breaking leading to bump states that move in response to integrated angular velocity.

Our work shows that large-scale mammalian neural circuits could implement continuous-attractor

dynamics using essentially the same mechanism as classical models—namely, through the instantiation of a continuous symmetry in the system—despite their heterogeneity and disorder. However, this symmetry is encoded more subtly than in classical models and may remain hidden in certain analyses of neuronal activity or connectivity. The symmetry is made manifest in the spectral structure of the connectivity and in the mean-field description of the network dynamics. Finally, we extend this framework to higher dimensions through an analysis of grid cells in the medial entorhinal cortex.

Our work relates to two recent studies pertaining to neuronal heterogeneity. While we were preparing this paper, Mainali et al. [26] reported the use of a similar Gaussian-process approach to characterize hippocampal place fields across different dimensionalities and environments. The success of Gaussian processes in both studies suggests their broad utility for describing heterogeneous neuronal responses. However, while Mainali et al. [26] model place-cell responses as arising from random feedforward connections from preexisting place-tuned inputs, we show that head-direction tuning can emerge through recurrent interactions.

Our approach also complements that of Darshan and Rivkind [27], who demonstrated that heterogeneous tuning curves can result from connectivity that combines a fixed random component and a learned structured component. We take a reverse, data-driven approach: rather than imposing constraints on network architecture, we derive the connectivity required to reproduce experimentally observed neuronal responses. This strategy departs from the common practice of optimizing networks to perform tasks. Instead, it directly fits network parameters to match experimental recordings at the single-neuron level, thus providing a general pipeline for connecting empirically observed neuronal tuning to circuit function [28, 29].

2 Examining a dataset of head-direction cells

We analyzed recordings of head-direction cells from the postsubiculum of freely moving mice [19] (data publicly available at [30]). Both the postsubiculum and the anterodorsal thalamic nucleus [18] contain neurons that encode head direction in a manner suggestive of ring-attractor dynamics, with activity patterns confined to a stable manifold with ring topology across different environments and during sleep [18, 19]. Previous analysis by Duszkiewicz et al. [19] emphasized the relative stereotypy of head-direction cells by contrasting them with more heterogeneous fast-spiking neurons. Nevertheless, head-direction cells themselves exhibit substantial heterogeneity in their responses, deviating markedly from the predictions of classical continuous-attractor models.

We analyzed $N_{\text{data}} = 1533$ excitatory neurons, pooled across 31 recording sessions, that were identified as head-direction cells by Duszkiewicz et al. [19]. These cells were classified based on encoding significant head-direction information relative to a time-reversed control distribution, without imposing constraints on the tuning-curve shapes such as unimodality. We constructed tuning curves by dividing head-direction angles into 100 bins and computing occupancy-normalized spike counts for each neuron. This procedure introduces high-frequency binning artifacts in the resulting tuning curves (Fig. 1D, “partition 1” and “partition 2”). To preserve genuine high-frequency features while reducing noise, we implemented a cross-validated smoothing procedure (SI.1). For each neuron, we divided the recording session into two partitions and computed separate tuning curves for each. We then smoothed

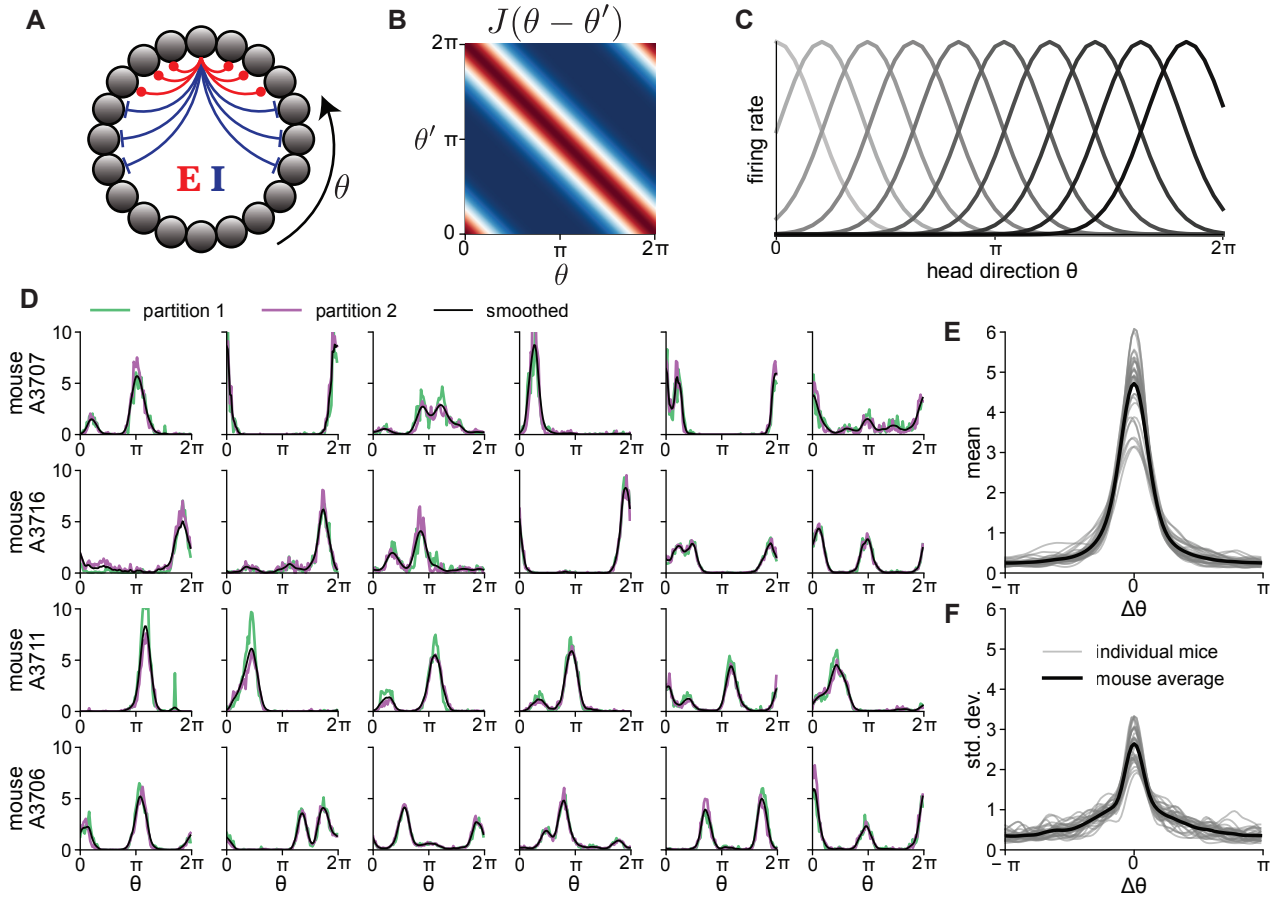

Figure 1: Classical continuous-attractor models fail to capture experimental data. (**A**) Schematic of a classical ring-attractor model with rotation-invariant connectivity. (**B**) Mexican-hat synaptic weight profile showing rotation-invariant connectivity of classical ring-attractor models. (**C**) Translation-invariant tuning curves predicted by classical ring-attractor models. Individual neuronal tuning curves have identical shapes, but their peaks are shifted to cover the θ axis uniformly. (**D**) Randomly chosen single-neuron tuning curves (arbitrary units due to normalization) from the dataset of [19, 30], selected from the four mice with the most recorded neurons. Green and purple traces show unsmoothed tuning curves computed from disjoint data partitions (SI.1). Black traces show cross-validated smoothed tuning curves. (**E**) Mean of center-of-mass-aligned tuning curves for each of the recording sessions (thin lines). Thick line represents the average across all sessions. (**F**) Standard deviation of center-of-mass-aligned tuning curves for each of 31 recording sessions (thin lines). Thick line represents the average across all sessions.

these curves using Gaussian kernels, selecting the optimal kernel width by maximizing the Poisson log-likelihood of spike counts in one partition given the smoothed tuning curve from the other partition. The final tuning curve for each neuron was obtained by averaging the two smoothed single-partition curves (Fig. 1D, "smoothed"). To facilitate comparison across neurons and animals, we normalized all tuning curves to have unit mean across head direction θ. These smoothed, normalized tuning curves form the basis for subsequent analyses.

2.1 Failure of classical ring-attractor models

Classical ring-attractor models (Fig. 1A, B) predict that all tuning curves within a circuit should have identical shapes up to neuron-specific shifts (Fig. 1C), reflecting the circular symmetry of the connectivity. However, examination of tuning curves from multiple recording sessions reveals significant heterogeneity, with neurons exhibiting variable numbers of irregularly spaced peaks (Fig. 1D). To quantify this heterogeneity, we aligned the tuning curves within each of the 31 recording sessions according to their circular centers of mass and computed both the mean tuning-curve profile (Fig. 1E)

and the fluctuations (standard deviation) around this profile (Fig. 1F). While the mean profiles were unimodal and symmetric, we observed substantial fluctuations comparable in magnitude to the means—a departure from the classical prediction of zero fluctuations. The angular profiles of both the means and fluctuations were remarkably consistent across sessions, despite the data spanning multiple animals. This consistency suggests a universal statistical structure underlying the observed heterogeneous responses, a feature we revisit in Sec. 3.

2.2 Network construction

The discrepancy between the observed tuning curves and the predictions of classical ring-attractor models motivated us to investigate whether a functional, quasi-continuous-attractor model could be constructed from the heterogeneous responses. While these responses could reflect feedforward input from a distinct attractor circuit elsewhere in the brain [31]—a possibility that, as is typical in neuroscience, cannot be definitively ruled out based on neural-activity data alone—we show that such an explanation is not necessary. Instead, we demonstrate that appropriate recurrent connections among the heterogeneous cells themselves could maintain this representation.

The model consists of a firing rate-based recurrent network of N neurons. While we use a rate-based formulation for analytical tractability, key behaviors of these networks generalize to their spiking counterparts [32]. The state of each neuron i, where $i = 1, \ldots, N$, is described by an input current $x_i(t)$ and firing rate $\phi_i(t)$. The network dynamics are governed by

$$\tau \partial_t x_i(t) = -x_i(t) + \sum_{j=1}^{N} J_{ij}\phi_j(t) + b_i, \tag{1}$$

where J_{ij} is the synaptic weight matrix, b_i is a vector of bias currents, and $\tau = 50$ ms is the network time constant. The term $-x_i(t)$ represents passive decay of neuronal activity, resulting in exponential relaxation to rest in the absence of input. The firing rate of a neuron is related to its input current x through a nonlinear function, $\phi(x)$. We choose a smooth and invertible nonlinearity to enable conversion between experimentally measured firing rates and the corresponding input currents. Specifically, we use a "softplus" function that interpolates between zero firing for strongly negative inputs and linear firing with unit gain for strongly positive inputs (see Sec. 3.2, Eq. 10; SI.2).

Using the dataset of normalized tuning curves pooled across mice, we defined a *target manifold* of activity patterns, $\phi_i^\star(\theta)$, where $i = 1, \ldots, N_{\text{data}}$. This manifold is embedded in N_{data}-dimensional neuronal space and parameterized by θ, with the periodicity of θ endowing it with the topology of a ring. To incorporate the target manifold into the network model, we first obtained the corresponding input-current manifold $x_i^\star(\theta)$ by applying the inverse of the nonlinearity to $\phi_i^\star(\theta)$. While these target patterns are *a priori* unrelated to the time-dependent network activities $\phi_i(t)$ and $x_i(t)$, our goal is to configure the synaptic weights J_{ij} and biases b_i such that network dynamics become confined to the target manifold. We formulated this as an optimization problem: at each point $x_i^\star(\theta)$ along the manifold, the time derivative should vanish: $\partial_t x_i(t) = 0$.

Due to the normalization of tuning curves (SI.2), the neuron-averaged firing rate is unity on the target manifold. As a result, the dynamics are invariant to a shift in the weights, $J_{ij} \to J_{ij} - c/N$, if

this is accompanied by a shift in the biases, $b_i \to b_i + c$. We exploit this property to prevent runaway activity due to the non-saturating nonlinearity, which is not necessarily prevented by $\partial_t x_i(t) = 0$ alone. In particular, we first optimize the weights J_{ij} with $b_i = 0$ and then select a sufficiently large $c > 0$ whose effect is to provide network-wide inhibition that stabilizes the mean activity mode.

We optimize the weights for the vanishing-flow condition using least-squares optimization, by minimizing the objective function

$$\mathcal{L}(\boldsymbol{J}; \boldsymbol{x}^\star) = \int_0^{2\pi} d\theta \sum_{i=1}^{N} \underbrace{\left(\sum_{j=1}^{N} J_{ij}\phi_j^\star(\theta) - x_i^\star(\theta) \right)}_{=\partial_t x_i(t) \text{ at point } \theta}^2 + \lambda \sum_{ij} J_{ij}^2, \tag{2}$$

subject to the constraint $J_{ii} = 0$ for all i. The first term in Eq. 2 ensures that, when network activity lies on the target manifold, its time derivative (given by Eq. 1) is small. It involves a sum over neurons and an integral over the manifold. The second term is a regularizer that prevents the weights from growing large (we use $\lambda = 10^{-6}$). Similar methodologies to fit networks directly to neural data have been used to study the goldfish oculomotor integrator [28], mouse posterior parietal cortex during a working-memory task [29], and zebrafish whole-brain activity [33]. The solution to this optimization problem can be expressed for $i \neq j$ as

$$J_{ij} = \frac{1}{N} \int_0^{2\pi} d\theta \int_0^{2\pi} d\theta' K_{i,\lambda}^{-1}(\theta, \theta') x_i^\star(\theta) \phi_j^\star(\theta'), \tag{3}$$

where the kernel $K_{i,\lambda}(\theta, \theta')$ is given by $K_{i,\lambda}(\theta, \theta') = C^\phi(\theta, \theta') - N^{-1}\phi_i(\theta)\phi_i(\theta') + \lambda\delta(\theta - \theta')$. Here, $C^\phi(\theta, \theta')$ is the two-point correlation function of the target firing rates,

$$C^\phi(\theta, \theta') = \frac{1}{N} \sum_i \phi_i^\star(\theta)\phi_i^\star(\theta'), \tag{4}$$

and the second term in the kernel accounts for the constraint $J_{ii} = 0$, which has a small effect for large N (SI.3).

The solution given by Eq. 3 has a structure reminiscent of Hebbian learning, with $\phi_j^\star(\theta')$ a presynaptic factor and $x_i^\star(\theta)$ a postsynaptic factor. Note that the nonlinearity appears only in the presynaptic term. Consequently, the weights do not form a symmetric matrix, $J_{ij} \neq J_{ji}$, and the resultant network behavior does not lend itself to a simple analysis of an energy function [34, 35].

Finally, to enable angular-velocity integration, a capability of biological head-direction systems, we incorporated angular-velocity inputs into the model. Classical ring-attractor models generally achieve this by adding an asymmetric component to the weights, derived from the angular derivative of the rotation-invariant connectivity profile [4]. We generalized this approach to the data-derived network model by adding an additional term to the connectivity, yielding

$$J_{ij}(\omega(t)) = \frac{1}{N} \int_0^{2\pi} d\theta \int_0^{2\pi} d\theta' K_{i,\lambda}^{-1}(\theta, \theta') \left(1 + \tau\omega(t)\partial_\theta\right) x_i^\star(\theta)\phi_j^\star(\theta'). \tag{5}$$

Here, $\omega(t)$ represents angular velocity. This expression can be derived by solving a least-squares optimization problem similar to Eq. 2, but with the objective of setting $\partial_t x_i(t) = \omega(t)\partial_\theta x_i^\star(\theta)$, where $\partial_\theta x_i^\star(\theta)$ is the tangent vector to the manifold at the point θ, instead of $\partial_t x_i(t) = 0$. While this approach nominally requires the full recurrent connectivity to be modulated by angular velocity, various biologically plausible mechanisms exist to mimic this effect [4, 9–11, 36, 37].

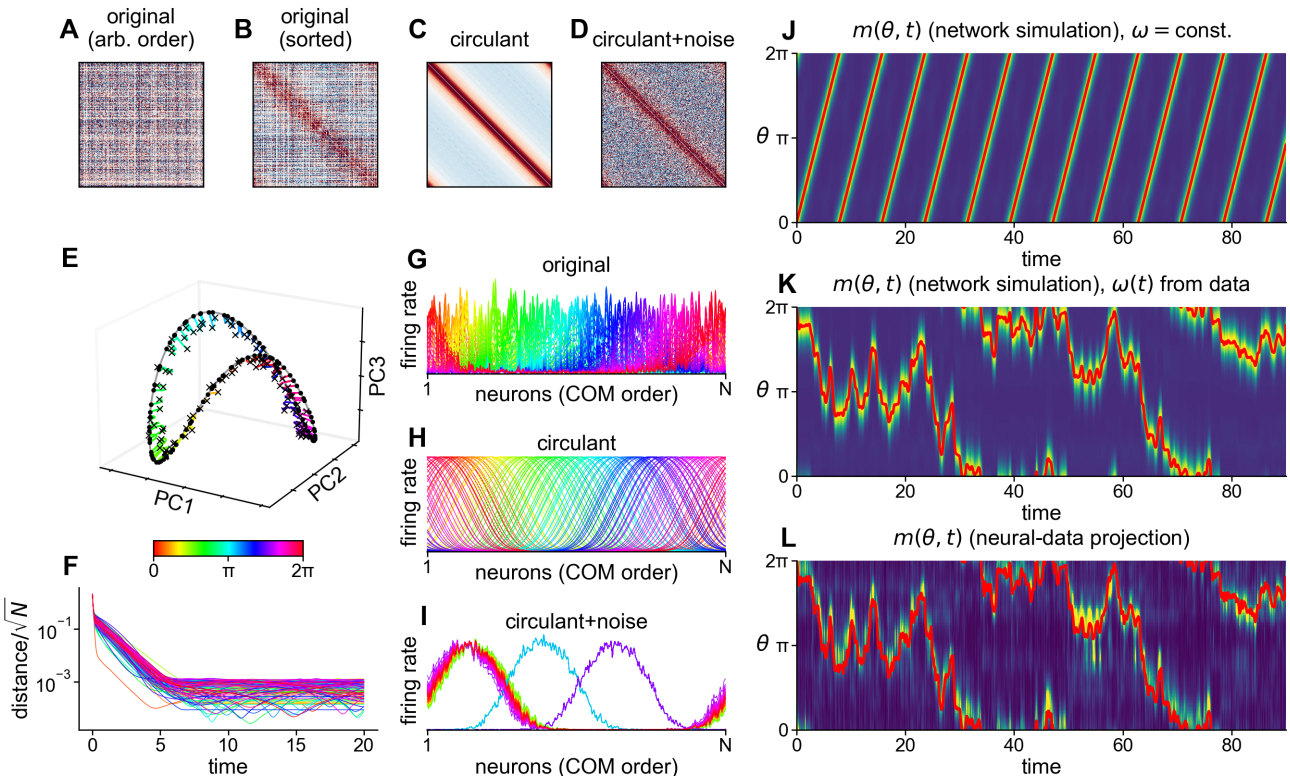

Figure 2: Quasi-continuous-attractor dynamics of the data-derived network. (**A**) Optimized weight matrix J_{ij} with neurons in random order, showing no apparent structure. (**B**) Same weight matrix as (A), but with neurons ordered according to the circular centers of mass of their tuning curves, revealing an underlying Mexican-hat-like structure. (**C**) Circularly symmetric version of matrix (B), obtained by averaging weights across diagonals corresponding to fixed angular differences between neurons. (**D**) Matrix from (C) with added noise, constructed by randomly permuting the residuals between matrices (B) and (C), demonstrating that random perturbation of the symmetric weights destroys the attractor properties. (**E**) Network dynamics visualized in the space of the first three principal components of the target manifold. Black x's mark initial conditions, colored lines show trajectories over 30 s of simulated time, and black dots indicate final states. In this and subsequent plots (G—J), colors indicate the angular locations of initial conditions. (**F**) Distance between each trajectory and its nearest point on the target manifold versus time, demonstrating convergence to the manifold. (**G**) Steady-state activity patterns across all neurons (ordered as in B) for different initial conditions, showing convergence to heterogeneous but stable bump configurations. (**H**) Steady-state patterns using the circularly symmetric matrix from (C), showing bump configurations equivalent up to translation, characteristic of classical ring-attractor models. (**I**) Steady-state patterns using the noise-corrupted matrix from (D), showing collapse to a small number of stable states with high-frequency noise, demonstrating that the specific structure of matrix (B) is essential for maintaining heterogeneous continuous-attractor dynamics. (**J**) Overlap order parameter $m(\theta, t)$ for constant angular-velocity input $\omega = 0.8$ rad/s, demonstrating smooth tracking of angular motion. (**K**) Overlap order parameter $m(\theta, t)$ computed from actual mouse head-direction trajectories, showing the network's ability to track natural motion. (**L**) Overlap order parameter $m(\theta, t)$ computed directly from mouse head-direction cell recordings, validating the model's prediction of population-level dynamics.

2.3 Dynamics of the data-derived network

We now examine the dynamics of the data-derived network, focusing on its function as a quasi-continuous ring attractor. We use uniform inhibition with strength $c = 1$ to stabilize the mean activity mode (for visualization purposes, we show the $c = 0$ weights in Fig. 2). As described in Sec.1, classical ring-attractor models rely on circular symmetry in connectivity to generate a continuous manifold of fixed points (Fig. 1A). Here, the constructed weights J_{ij} do not exhibit this symmetry in an obvious way. When neurons are arranged in arbitrary order, the weight matrix J_{ij} appears, necessarily, unstructured (Fig. 2A). However, when neurons are arranged according to their circular centers of mass, an approximate Mexican-hat structure is revealed (Fig. 2B).

For sufficiently large N, the optimization problem used to construct the weights is overparameterized: each row has more degrees of freedom (N) than effective constraints (roughly the number of large Fourier components in the tuning curves). Consequently, the optimized weights achieve very small values of $\partial_t x_i(t)$ everywhere on the target manifold. However, this alone does not guarantee attractor stability, which requires network activity to remain confined to this manifold and return to it when perturbed. The uniform inhibition prevents the divergence of the mean activity mode but does not guarantee confinement to the target manifold. To test the attractor properties of the network, we initialized the network at states separated by random noise from points uniformly spaced on the target manifold and observed the dynamics over 30 s of simulation time. All initial conditions converge rapidly ($\sim$1 s) to nearby points on the target manifold (Fig. 2E, F). This stability likely emerges from the small regularization term in the loss function that, as $\lambda \to 0$, selects the minimum-norm solution in the overparameterized regime. Small weights promote stability by ensuring that most eigenvalues of the Jacobian matrix, evaluated on the target manifold, are substantially negative, corresponding to rapid relaxation back to this manifold. This relationship between overparameterization and dynamical stability has been noted in previous work [38]. We demonstrate stability rigorously in the large-network limit using dynamical mean-field theory in Sec. 4.2.

To probe what structural features of the optimized weights are required to allow for both heterogeneous responses and quasi-continuous-attractor behavior, we compare the data-derived weight matrix to two variants. First, we create a circularly symmetric version by averaging weights over off-diagonal bands corresponding to fixed angular differences between centers of mass (Fig. 2C). Second, we randomly permute the residuals between the data-derived and circularly symmetric matrices, randomizing the non-circulant component (Fig. 2D). For both constructed matrices, we amplified the strengths of all weights by a multiplicative factor to produce bump formation rather than collapse to a homogeneous state (SI.4). Examining steady-state activity patterns across neurons, ordered by their center-of-mass angles, demonstrates the importance of the specific weight-matrix structure. As we have shown, the data-derived weights maintain a quasi-continuum of stable, heterogeneous states (Fig. 2G), and, as expected, the circularly symmetric weights produce translation-invariant bump configurations characteristic of classical ring attractors (Fig. 2H). In contrast, the noise-corrupted weights cause the network to collapse onto a small number of discrete stable states whose tuning curves are noisy versions of those of the circulant system (Fig. 2I). Thus, the specific structure of heterogeneity in the data-derived weights is essential for maintaining quasi-continuous-attractor dynamics that reproduce the observed neuronal responses.

To characterize the alignment between network firing rates $\phi_i(t)$ and the target manifold $\phi_i^\star(\theta)$, we define an *overlap order parameter*, given by

$$m(\theta, t) = \frac{1}{N} \sum_{i=1}^{N} \phi_i^\star(\theta) \phi_i(t). \tag{6}$$

When network activity lies exactly on the ring at position ψ, we have $\phi_i(t) = \phi_i^\star(\psi)$, resulting in $m(\theta, t) = C^\phi(\theta, \psi)$ (Eq. 4). As a function of θ, this produces a localized bump of activity peaked at ψ. This situation parallels classical ring-attractor models, which are similarly described by a time-dependent order parameter that exhibits steady-state localized bumps. Similarly, when the network state moves along the target manifold in response to angular-velocity input, we expect $m(\theta, t)$ to take the

form of a traveling bump of activity centered around the integrated angular displacement.

Using $m(\theta, t)$ as a readout of network activity, we tested the network's ability to integrate angular-velocity inputs. For a constant input $\omega = 0.8$ rad/s, visualization of $m(\theta, t)$ shows accurate tracking (Fig. 2J). The network also successfully tracks more complicated trajectories derived from the mouse head-direction data (Fig. 2K), demonstrating robust integration of realistic velocity inputs.

The overlap order parameter $m(\theta, t)$ can also be computed directly from neural recordings by projecting measured firing rates onto measured tuning curves. Using the definition in Eq. 6, $\phi_i^\star(\theta)$ is still the measured tuning curve for neuron i, but we now take $\phi_i(t)$ to be the binned spike counts of neuronal time-series data. The resulting bump pattern accurately tracks the animal's head direction (Fig. 2L) and closely resembles the patterns generated by the network model (Fig. 2K). This correspondence provides strong evidence that neuronal activity remains confined to a manifold closely aligned with that defined by the tuning curves [18]. Furthermore, the peak of $m(\theta, t)$ serves as an effective decoder of head direction—in fact, it provides the maximum *a posteriori* estimate of head direction under the assumption of Gaussian noise around the tuned neuronal responses and a uniform prior over θ.

3 Statistical generative process

While our analysis suggests that continuous-attractor dynamics can emerge from heterogeneous neuronal responses, we face limitations due to dataset size when using experimental recordings. The experimental data include $N_{\text{data}} = \mathcal{O}(10^3)$ neurons when pooled across mice, but individual recording sessions typically involve fewer neurons, at most $\mathcal{O}(10^2)$. Such session-specific analyses are necessary for studying the distribution of tuning responses because phases cannot be meaningfully compared between neurons recorded in different animals. Even the full sample sizes represent only a small fraction of the head-direction cell population in the postsubiculum—a region likely containing tens to hundreds of thousands of such neurons. To overcome this limitation and extend our analysis to scales characteristic of mammalian neural circuits, we developed a statistical generative process capable of producing large ensembles of tuning curves. This approach allows us to study network properties in the large-N limit while faithfully preserving the structure of the data.

3.1 Evidence for distributional circular symmetry

Even when individual tuning curves appear heterogeneous, they may arise from an underlying generative process with circular symmetry. Such a symmetry would have two important implications. First, it would reveal a simplification or "inductive bias" in the generative process itself. Second, it would suggest a mechanism through which circular symmetry could emerge at the population level, despite the disorder present in single-neuron responses.

To test this possibility, we first examined the distribution of center-of-mass angles of tuning curves within each recording session, noting that relative phase relationships cannot be meaningfully compared between neurons recorded in different animals (Fig. 3A). Visual inspection suggests that these distributions could be uniform (Fig. 3B), which we confirmed using the Kuiper test for circular uniformity, followed by a Benjamini-Hochberg correction for multiple comparisons across sessions. This revealed

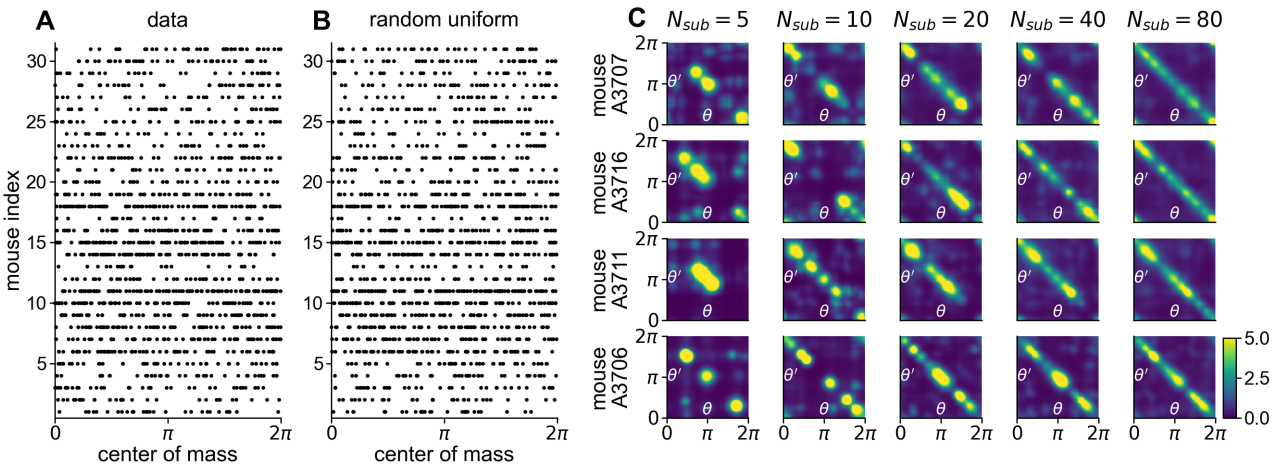

Figure 3: Evidence of distributional circular symmetry in head-direction cell responses. (**A**) Distribution of preferred angles across neurons. Each row represents a different recording session, with dots indicating the center-of-mass angles of individual tuning curves. (**B**) Random angle distributions matched to the number of neurons in each session, demonstrating visual similarity to the experimental data in (A). Statistical analysis confirms that the experimental distributions are consistent with uniformity. (**C**) Empirical two-point correlation functions, $C^{\phi}(\theta, \theta')$, computed from three example recording sessions, shown for increasing subset sizes (N_{subset}). The correlation structure becomes increasingly circulant as more neurons are included, revealing a distributional circular symmetry.

no statistically significant deviation from uniformity in any session, indicating an absence of systematic bias in center-of-mass angles that provides one signature of an underlying circular symmetry in the distribution of tuning curves.

As a stronger test of this distributional symmetry, we examined the two-point correlation function of tuning curves, $C^{\phi}(\theta, \theta')$ (Eq. 4). When computed for increasingly large random subsets of neurons, this function appears to approach a circulant form (Fig. 3C), consistent with circular symmetry in the tuning-curve distribution.

3.2 Structure of the generative process

Motivated by these observations, we developed a two-step generative process to sample realistic tuning curves. The first step involves sampling an input-current tuning curve $x^{\star}(\theta)$ from a circularly symmetric Gaussian process with zero mean and covariance function $\Gamma^{x}(\Delta\theta)$,

$$\langle x^{\star}(\theta) \rangle = 0, \tag{7}$$

$$\langle x^{\star}(\theta) x^{\star}(\theta') \rangle = \Gamma^{x}(\theta - \theta'), \tag{8}$$

where the covariance function depends solely on the angular difference $\Delta\theta = \theta - \theta'$ in accordance with the evidence for distributional circular symmetry presented above. It is parameterized as

$$\Gamma^{x}(\Delta\theta) = \sum_{n=-\infty}^{\infty} \exp\left(-\frac{\sigma^2}{2}(\Delta\theta + 2\pi n)^2\right), \tag{9}$$

where σ controls the typical number of peaks in the generated tuning curves, or equivalently, the number of large Fourier components in these tuning curves (Fig. 4A, B). The second step transforms these input currents into firing rates using a normalized softplus function,

$$\phi^{\star}(\theta) = \frac{1}{Z[x^{\star}]} \log\left(1 + e^{\beta(x^{\star}(\theta) - b)}\right), \tag{10}$$

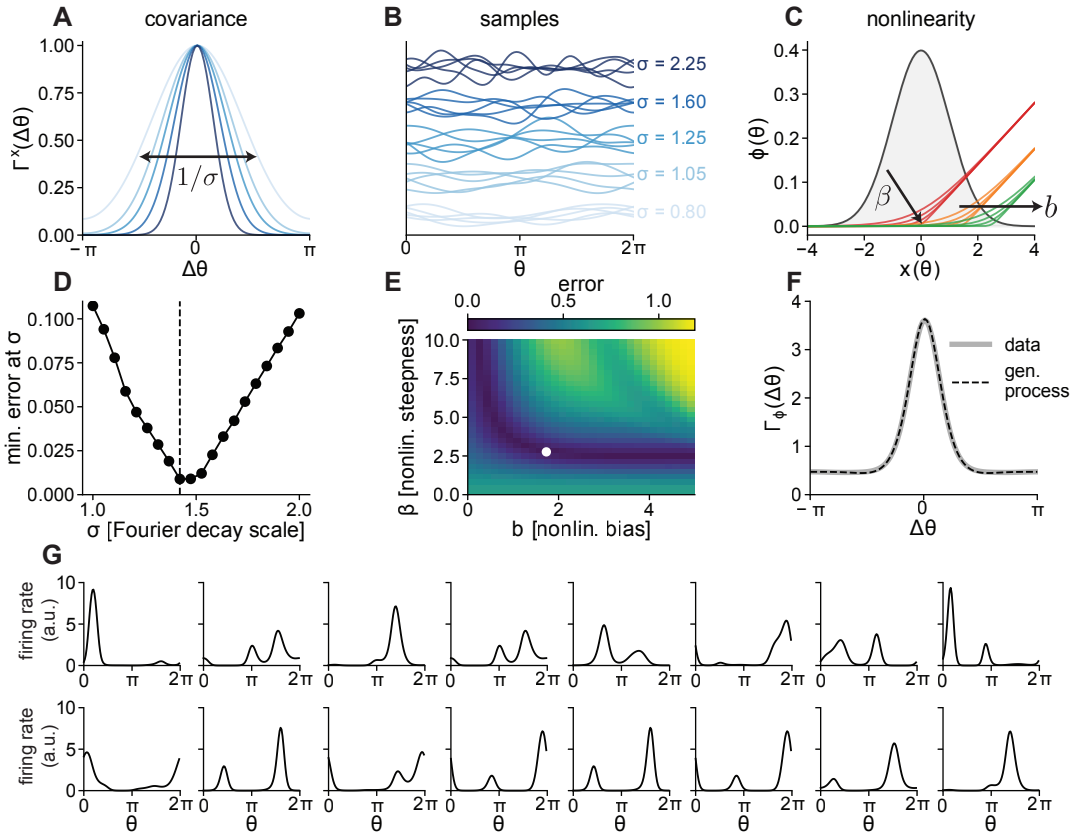

Figure 4: Illustration of the statistical generative process. (**A**) Angular covariance functions $\Gamma^x(\Delta\theta)$ for different values of the Fourier decay scale σ, showing how this parameter controls the smoothness of generated tuning curves. (**B**) Example samples from the Gaussian process prior for different values of σ, demonstrating the relationship between σ and tuning curve complexity. (**C**) Effect of nonlinearity parameters β (sharpness) and b (threshold) on the input-output relationship. Gray shading shows the Gaussian distribution of input currents. (**D**) Model selection curve showing the minimal error (optimized over b and β) as a function of σ. Error measures the difference between model and experimental two-point correlation functions. (**E**) Error landscape at optimal σ as a function of b and β, revealing a clear minimum in parameter space. (**F**) Two-point correlation functions of the data and generative process agree at the optimal parameter values. (**G**) Random samples from the optimized generative process, exhibiting the complex, heterogeneous structure characteristic of experimental tuning curves.

where β controls the sharpness of the nonlinearity, b sets a soft threshold, and $Z[x^\star]$ is a functional that normalizes the mean over θ of the firing rate to unity (Fig. 4C). This normalization matches our processing of the experimental data.

3.3 Quantitative validation

Rather than fitting session-specific parameters, we sought a single parameter set (σ, β, and b) for the generative process that could capture the statistical structure across all recordings. We fit these parameters by minimizing the difference between the empirical two-point correlation function, pooled across recording sessions, and the generative correlation function. (Fig. 4D–F). The optimized generative process yields tuning curves that closely resemble the experimental observations (Fig. 4G).

To quantitatively validate the generative process, we compared several statistical features of generated and experimental tuning curves (Fig. 5). First, we computed the center-of-mass-aligned mean tuning-curve profile, as well as the profile of fluctuations, for the generative process (as done for the data in Sec. 2.1). These predicted profiles match those observed experimentally (Fig. 5A, B).

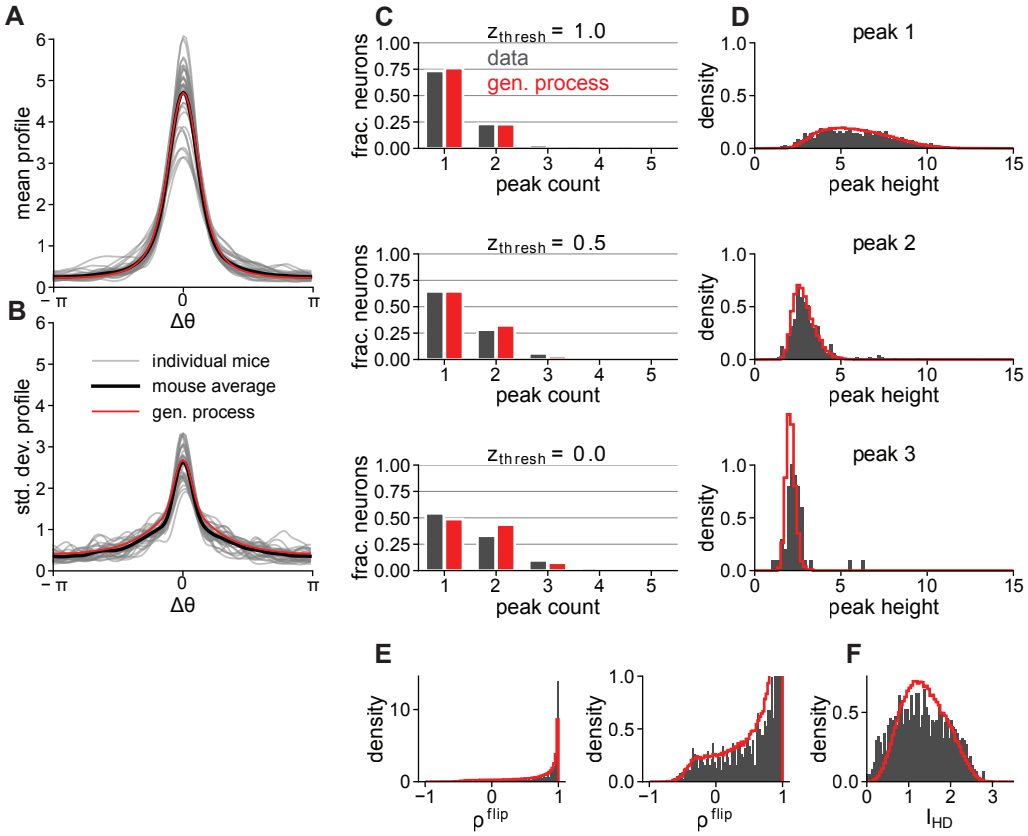

Figure 5: Detailed statistical comparison between model-generated and experimental tuning curves. (**A**) Mean tuning-curve profile predicted by the generative process (red) overlaid on experimental data from Fig. 1E (individual sessions in gray, session average in black). (**B**) Standard deviation of tuning curves predicted by the generative process (red) compared to experimental measurements from Fig. 1F (individual sessions in gray, session average in black). (**C**) Distribution of peak counts per tuning curve for decreasing detection thresholds (z_{thresh}, top to bottom). Here and in the following panels, Blue: experimental data; black: model. (**D**) Distributions of tuning curve values at the three highest local maxima. The model accurately captures both the central tendency and variability of peak heights. (**E**) Distribution of flip symmetry correlations (ρ^{flip}), with an expanded view of the asymmetric tail (right). The model reproduces the full range of tuning curve symmetries observed experimentally. (**F**) Distribution of head-direction information content (I_{HD}), showing agreement between model and data across the full range of coding precision.

We next analyzed the distribution of peak counts in individual tuning curves, where a peak is counted only if its z-score (computed using the tuning curve's own mean and standard deviation) exceeds a detection threshold z_{thresh} (Fig. 5C). The model accurately reproduces the experimental peak-count distributions, with particularly strong agreement at higher detection thresholds z_{thresh}.

We further characterized the peak structure by comparing the full distributions of tuning-curve values at the first, second, and third highest peaks, using $z_{\text{thresh}} = 1$ (Fig. 5D). Again, we find close agreement, now for full distributions, between the generative process and the data.

Going beyond peak structure, we examined the deviation from reflection symmetry in individual tuning curves. For each tuning curve, we computed the correlation ρ^{flip} between the original tuning curve and its reflection around its circular center of mass. The model accurately reproduces the distribution of these correlations (Fig. 5E, left), capturing particularly well the tail of strongly asymmetric tuning curves (Fig. 5E, right).

Finally, we compared the distributions of head-direction information content I_{HD}, a commonly used measure of how precisely individual neurons encode heading [39] (Fig. 5F). The model shows good

agreement with the data, especially for the tail of neurons with high information content.

In summary, the generative process accurately reproduces an array of features of the experimental tuning curves, indicating that the parsimonious three-parameter model has identified universal statistical structure present in the rodent head-direction system. The particularly strong agreement in the tails of these distributions suggests possible universal extreme value statistics at play, though we note that agreement is seen for non-extreme quantities as well. On the basis of the success of the generative process, we use these synthetic tuning curves as simulated experimental data in our optimization-based construction of recurrent-network models, the topic of the next section.

4 Large-network limit and dynamical mean-field theory

Using tuning curves sampled from the generative process, we study the properties of target manifolds, optimized connectivity matrices, and network dynamics at scales reflecting mammalian cortical circuits. Specifically, we take $N \to \infty$.

For analytical tractability in this and subsequent sections, we adopt the saturating firing rate nonlinearity $\phi(x) = 1 + \mathrm{erf}(\beta\sqrt{\pi}x/2)$, where β controls the sharpness of the response. This function saturates at 0 and 1 for large negative and positive inputs, respectively, allowing us to set the stabilizing term c in the synaptic weights to zero.

4.1 Large-N analysis of optimized connectivity

The target manifold consists of the desired patterns of input currents $x_i^\star(\theta)$ and firing rates $\phi_i^\star(\theta)$ across neurons. We organize these patterns into matrices $\boldsymbol{X}$ and $\boldsymbol{\Phi}$ of size $N_\theta \times N$, where we have discretized θ. Although the tuning curves may contain infinitely many Fourier components, due to their smoothness we can truncate them at a frequency D that does not scale with N while introducing negligible error. These matrices consequently have rank $2D + 1$, with singular value decompositions

$$\boldsymbol{X} = \boldsymbol{U}^x \mathrm{diag}(\boldsymbol{S}^x)(\boldsymbol{V}^x)^T \text{ and } \boldsymbol{\Phi} = \boldsymbol{U}^\phi \mathrm{diag}(\boldsymbol{S}^\phi)(\boldsymbol{V}^\phi)^T, \tag{11a}$$

where $\boldsymbol{S}^x = (S_0^x, S_1^x, \ldots, S_{2D}^x)$ and $\boldsymbol{S}^\phi = (S_0^\phi, S_1^\phi, \ldots, S_{2D}^\phi)$ are vectors of singular values. The matrices $\boldsymbol{U}^x$, $\boldsymbol{U}^\phi$, $\boldsymbol{V}^x$, and $\boldsymbol{V}^\phi$, each of size $N \times (2D + 1)$, contain left and right singular vectors with orthonormal columns. As $N \to \infty$, the distributional circular symmetry identified in the data (Sec. 3.1) and incorporated into the generative process (Sec. 3.2) becomes a geometric circular symmetry of these manifolds. Specifically, the left singular vectors converge to Fourier modes, $\boldsymbol{U}^x = \boldsymbol{U}^\phi = \boldsymbol{F}$, where $\boldsymbol{F}$ contains the constant mode followed by sine-cosine pairs for each frequency up to D. The matrices $\boldsymbol{V}^x$ and $\boldsymbol{V}^\phi$ describe how these modes are orthogonally embedded in the high-dimensional spaces of input currents and firing rates. The singular values, which set the amplitudes of different harmonics, converge to the square roots of the Fourier coefficients of their respective correlation functions, $\hat{\boldsymbol{\Gamma}}^x = (\hat{\Gamma}_0^x, \hat{\Gamma}_1^x, \ldots, \hat{\Gamma}_{2D}^x)$ and $\hat{\boldsymbol{\Gamma}}^\phi = (\hat{\Gamma}_0^\phi, \hat{\Gamma}_1^\phi, \ldots, \hat{\Gamma}_{2D}^\phi)$, with consecutive pairs (excluding the constant mode) being equal, exhibiting a doublet degeneracy.

The optimized weight matrix $\boldsymbol{J}$ connects these two target manifolds. Writing the solution for $\boldsymbol{J}$ (Eq. 3) in terms of the singular value decompositions above and using the large-N relation $(\boldsymbol{U}^x)^T\boldsymbol{U}^\phi =$

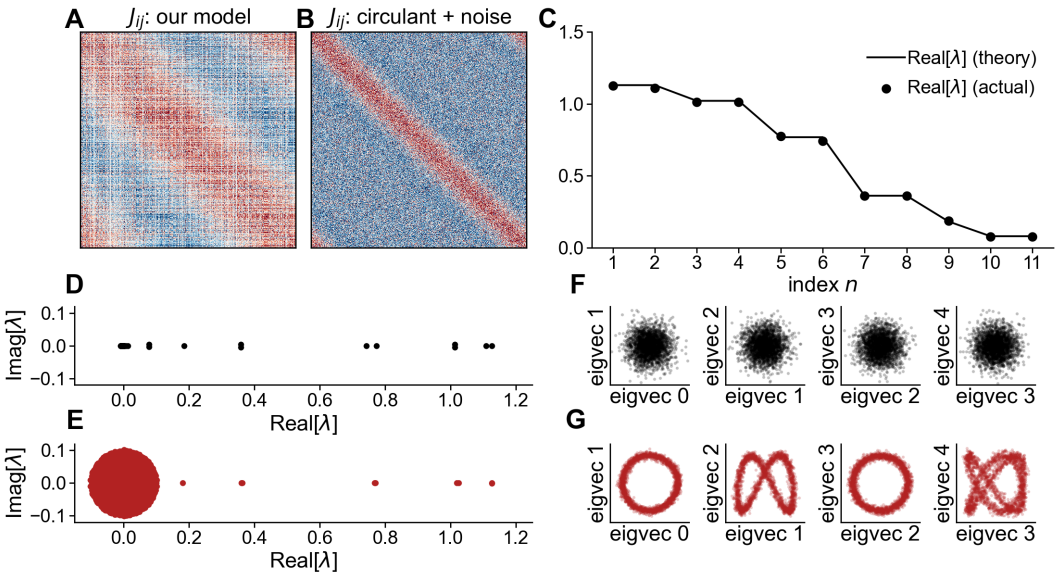

Figure 6: Spectral structure of the optimized weight matrix. (**A**) Optimized weight matrix using tuning curves drawn from the generative process ($N = 2500$), with σ and β values from the fitted model. Neurons are ordered by their tuning-curve centers of mass. (**B**) Circulant matrix ($N = 2500$) with added noise, where the circulant component is scaled by $1/N$ and the noise term is of order $1/\sqrt{N}$. (**C**) Sorted real parts of the eigenvalues of the matrix from (A). Theoretical predictions (lines) are shown alongside numerical results (dots), with visible doublet degeneracy. Mode 9 is the constant mode and thus does not occur in a doublet. (**D**) Complex-plane view of eigenvalues of the optimized weight matrix, with real parts as in (C). (**E**) Complex-plane view of eigenvalues from the circulant-plus-noise matrix shown in (B). (**F**) Scatterplots comparing consecutive pairs of eigenvectors of $\boldsymbol{J}\boldsymbol{J}^T$ (equivalently, the left singular vectors of $\boldsymbol{J}$) where $\boldsymbol{J}$ is the matrix from (A). The absence of structure reflects the disordered embedding. (**G**) Analogous scatterplots for the circulant-plus-noise matrix, where the leading eigenvectors retain their Fourier structure despite partial corruption by noise, revealing an underlying ring structure.

$\boldsymbol{F}^T\boldsymbol{F} = \boldsymbol{I}_{2D+1}$ implies that

$$\boldsymbol{J} = \boldsymbol{V}^x \mathrm{diag}\left(\sqrt{\frac{\hat{\boldsymbol{\Gamma}}^x}{\hat{\boldsymbol{\Gamma}}^\phi}}\right)(\boldsymbol{V}^\phi)^T, \tag{12}$$

where we have taken $\lambda \to 0$. The same large-N relation implies that the singular-vector overlaps are diagonal: $(\boldsymbol{V}^\phi)^T\boldsymbol{V}^x = \alpha\,\mathrm{diag}(\sqrt{\hat{\boldsymbol{\Gamma}}^x/\hat{\boldsymbol{\Gamma}}^\phi})$, where $\alpha = \langle\phi'(\theta)\rangle$ is the average slope of the neuronal activation function. Consequently, $\boldsymbol{V}^x$ and $\boldsymbol{V}\phi$ comprise the right and left eigenvectors of $\boldsymbol{J}$ with eigenvalues

$$\lambda_n = \alpha\hat{J}_n, \text{ where } \hat{J}_n = \frac{\hat{\Gamma}^x_n}{\hat{\Gamma}^\phi_n}. \tag{13}$$

These eigenvalues are real and exhibit doublet degeneracy, reflecting an underlying circular symmetry (Fig. 6A, C, and D). In particular, Eq. 13 suggests that the dynamical behavior of the system is related to that of a *classical* ring-attractor model with symmetric and rotation-invariant connectivity proportional to $J(\Delta\theta)$, given by the inverse Fourier transform of $\hat{J}_n$. We establish this connection rigorously through dynamical mean-field theory in Sec. 4.2.

Our disordered model realizes its circular symmetry differently from classical ring-attractor models. In classical models with circulant connectivity, the eigenvectors of $\boldsymbol{J}$, and likewise the matrices that embed the ring manifold into the spaces of input currents and firing rates, are Fourier modes over the ring of neurons—in our notation $\boldsymbol{V}^\phi = \boldsymbol{V}^x = \boldsymbol{F}$. In contrast, our model's circular symmetry is "hidden" since $\boldsymbol{V}^x$ and $\boldsymbol{V}^\phi$ are disordered.

This disorder has important implications for connectome analysis. Consider the challenge of identifying ring structures in an observed weight matrix. A common method is spectral embedding, which involves computing the top few eigenvectors of matrices derived from the connectivity. These could be eigenvectors of the Laplacian or, as shown in Fig. 6F, of $\boldsymbol{J}\boldsymbol{J}^T$—the latter being the left singular vectors of $\boldsymbol{J}$. These eigenvectors are then used to assign coordinates to each neuron, with the resulting coordinate structure visualized to detect underlying geometric patterns. Although this method successfully reveals rings in circulant matrices, where the eigenvectors are Fourier modes, it fails for our model because our eigenvectors are random (Fig. 6F).

A more reliable signature of ring structure lies in the eigenvalue spectrum rather than in the eigenvector structure, which is sensitive to how the ring is embedded in neuronal space. Two key features of the eigenvalues are particularly notable. First, they are real, a surprising property given that $\boldsymbol{J}$ maintains a finite level of asymmetry even as $N \to \infty$, characterized by $\mathrm{corr}(J_{ij}, J_{ji}) = \alpha^2[J^2](0)/J(0)$. Second, and perhaps most significantly, the eigenvalues exhibit doublet degeneracy, providing a robust signature of the underlying circular symmetry in the disordered model.

A natural alternative model to consider is one in which independent random noise is added to a circulant connectivity matrix. In this case, unlike our model, the spectral embedding approach remains viable for detecting ring structure as long as the $1/\sqrt{N}$ noise is not too strong. While noise affects the eigenvalue spectrum by adding a bulk distribution (Fig. 6E), the eigenvectors retain their Fourier structure until the noise becomes strong enough for the bulk eigenvalues to overwhelm the outliers (Fig. 6G). This is a fundamental difference from our model, where the eigenvectors are random by construction. When the noise becomes too strong, the ring attractor becomes dysfunctional even as $N \to \infty$ due to the emergence of chaotic dynamics. Meanwhile, small amounts of noise reduce the system to a small number of quasi-stable states at finite-N (Fig. 2I), indicating that the disordered model is much more robust. In both models, a continuous ring of fixed points is recovered as $N \to \infty$.

This distinction between our model and the circulant-plus-noise model manifests in the mean-field description of the network dynamics, studied below. Recurrent neural networks with disorder generally fall into two classes: those with order-$1/N$ connectivity and low-rank weights, and those with order-$1/\sqrt{N}$ connectivity and extensive-rank weights (where "extensive" means proportional to N). The order parameters are time-dependent overlaps in the former case and correlation functions describing fluctuations in the latter. Our system belongs to the first class, while the circulant-plus-noise case combines both types, requiring tracking of high-dimensional fluctuations absent in our model.

In summary, our analysis shows that ring attractors can be implemented through either ordered Fourier embeddings (as in classical models) or disordered random embeddings (as in our model). The strategy that biological circuits use could be determined through analysis of future connectome datasets [40], where an important signature will be whether the ring structure appears in the eigenvectors or only in the eigenvalues.

To understand how the spectral properties elucidated here shape the network's nonlinear dynamics, we now develop a dynamical mean-field theory in the large-N limit.

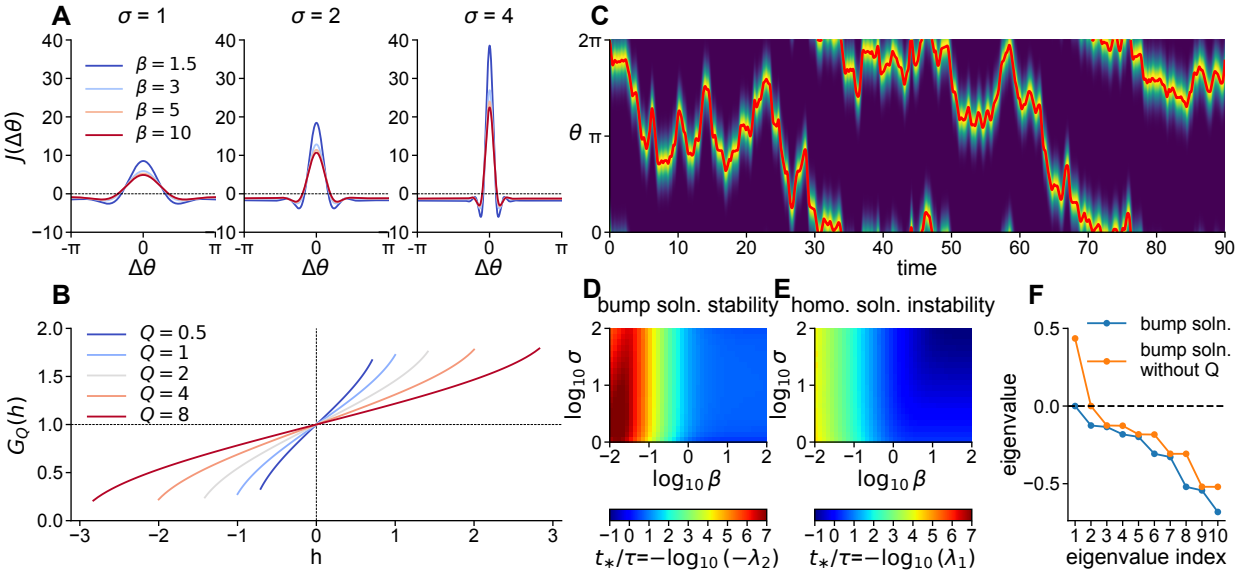

Figure 7: Dynamical mean-field theory reveals circular symmetry and ring-attractor dynamics in the large-network limit. (**A**) Mean-field weights $J(\Delta\theta)$ exhibit a Mexican-hat profile that varies with the generative process parameters σ (panels) and nonlinearity parameter β (curves). Higher β increases the amplitude of excitation and inhibition, while larger σ narrows the profile. (**B**) Effective nonlinearity $G_Q(h)$ for different values of the activity-dependent gain parameter Q. As network activity increases, Q grows, reducing the slope of the nonlinearity and stabilizing bump formation through an emergent gain-control mechanism. (**C**) Solution of the dynamical mean-field equations showing a traveling bump tracking head-direction velocity input, matching the dynamics observed in the high-dimensional network. (**D**) Timescale of convergence to the ring manifold ($t_*/\tau = -\log(-\lambda_2)$) as a function of σ and β, showing rapid convergence ($t_* \sim \tau$) for $\beta \gtrsim 1$ across all σ values. (**E**) Timescale of divergence from the uniform state ($t_*/\tau = -\log(\lambda_1)$), demonstrating pattern-forming (Turing) instability with similar parameter dependence. (**F**) Real parts of the leading eigenvalues at the bump state for the full dynamical mean-field equations (blue) and modified equations with fixed $Q = \Gamma^x(0)$ (orange). Dynamic $Q(t)$ is essential for bump stability, as fixing $Q(t)$ introduces a positive eigenvalue that destabilizes the bump.

4.2 Dynamical mean-field theory

In this section, we go beyond our prior analysis by developing a theory of the full network dynamics in the limit $N \to \infty$. Strikingly, a dynamical mean-field theory recovers all essential features of classical ring-attractor models: Mexican-hat connectivity, nonlinear pattern formation, and continuous circular symmetry. We present the main components here, with details given in SI.5.

The dynamical mean-field theory reduces the high-dimensional network dynamics to equations governing the dynamics of a low-dimensional order parameter, $\tilde{m}(\theta, t)$, which is a temporally low-pass filtered version of the overlap order parameter $m(\theta, t)$ defined in Eq. 6. Specifically, for heading movements with small angular acceleration ($\tau^2|\partial_t\omega(t)| \ll 1$, which is satisfied by the behavioral conditions), this order parameter obeys

$$\partial_t \tilde{m}(\theta, t) = -\tilde{m}(\theta, t) + G_{Q(t)}\left(\int_0^{2\pi} d\theta' \left[1 - \tau\omega(t)\partial_\theta\right] J(\theta - \theta')\tilde{m}(\theta', t)\right). \tag{14}$$

This equation is very similar to that of classical ring-attractor models (see, e.g. [2]), although it has a more complex structure in the nonlinearity $G_{Q(t)}(h)$, which depends on a dynamic effective gain variable $Q(t)$ in addition to its input h, as discussed below.

The effective interactions appearing in Eq. 14 are rotation-invariant weights $J(\Delta\theta)$ defined in Fourier space as $J_n = \Gamma^x_n/\Gamma^\phi_n$, directly related to the eigenvalues of $\boldsymbol{J}$ (Eq. 13). These effective

weights, visualized in Fig. 7A, differ markedly from the full network weights J_{ij}. While J_{ij} lacks both reciprocal symmetry ($J_{ij} \neq J_{ji}$) and a simple form of rotation symmetry, $J(\Delta\theta)$ is symmetric ($J(\Delta\theta) = J(-\Delta\theta)$) and explicitly rotation-invariant. The weights $J(\Delta\theta)$ exhibit a Mexican-hat profile characteristic of pattern-forming systems, with the shape controlled by the generative-process parameter σ and nonlinearity parameter β (Fig. 7A). Increasing β amplifies both excitation and inhibition, while larger σ produces a narrower profile.

The nonlinearity $G_{Q(t)}(h)$ in Eq. 14 involves a factor $Q(t)$ (Fig. 7B) that depends on the instantaneous angular profile of $\tilde{m}(\theta, t)$ at time t and is a measure of the level of global excitation across angles θ. Specifically, $Q(t)$ is a quadratic functional of the form $Q(t) = \int_0^{2\pi} d\theta \int_0^{2\pi} d\theta' U_{\omega(t)}(\theta - \theta')\tilde{m}(\theta, t)\tilde{m}(\theta', t)$, where $U_{\omega(t)}(\Delta\theta)$ is an angular velocity-dependent kernel (SI.5). Unlike the dynamical equations of classical ring-attractor models and of the disordered N-dimensional network model, both of which use local, pointwise nonlinearities, the function $G_{Q(t)}(h)$ is shaped by global system activity through $Q(t)$. In particular, as activity increases and $Q(t)$ increases, the slope of $G_{Q(t)}(h)$ decreases. Thus, $Q(t)$ acts as a state-dependent gain-control parameter, absent in the original single-neuron dynamics. This state dependence is required for the stability of bump solutions, as demonstrated below.

We now discuss solutions of the dynamical mean-field equations and their stability. For $\omega(t) = 0$ the equations exhibit static solutions $\tilde{m}(\theta) = m(\theta)$. The stable solutions are bump states of the form

$$m(\theta) = \Gamma^{\phi}(\theta - \psi), \tag{15}$$

where ψ sets the angular location of the peak of the bump, reflecting the rotation-invariance of the dynamical mean-field equations. For $\omega(t) \neq 0$, there are traveling-bump solutions,

$$m(\theta, t) = \Gamma^{\phi}\left(\theta - \psi - \int^{t} dt' \omega(t')\right), \tag{16}$$

with $\tilde{m}(\theta, t)$ following a temporally smoothed version of this profile. These solutions track head direction through integration of angular velocity (Fig. 7C), matching the behavior observed in both network simulations (Fig. 2K, L) and experimental recordings (Fig. 2M). We note that, for both static and traveling-bump solutions, $Q(t) = \Gamma^{x}(0)$, independent of time.

The stability properties and related timescales of the mean-field dynamics can be analyzed through the Jacobian $M(\theta, \theta') = \partial_{\tilde{m}(\theta', t)}\left[\partial_t \tilde{m}(\theta, t)\right]$ (SI.6). When evaluated at a particular equilibrium state $\tilde{m}(\theta)$, the absence or presence of an eigenvalue with a positive real part indicates stability or instability, respectively. We always expect a single zero eigenvalue due to the continuous symmetry of the system.

For a bump state $\tilde{m}(\theta) = \Gamma^{\phi}(\theta - \psi)$ (where ψ is arbitrary and can be set to $\psi = 0$), diagonalization of $M(\theta, \theta')$ reveals that all eigenvalues have negative real parts except for a single zero eigenvalue corresponding to rotation symmetry (Fig. 7F, blue curve). The ultimate convergence timescale to the ring manifold is determined by the eigenvalue λ_2 with the second-largest real part, which is itself real. Specifically, perturbations decay with a characteristic timescale $t_*/\tau = -\log(-\lambda_2)$ (Fig. 7D). The timescale t_* is of order τ (i.e., convergence is fast) for $\beta \gtrsim 1$.

The system exhibits a Turing instability that destabilizes the uniform state $\tilde{m}(\theta) = \tilde{m}_0$, where $\tilde{m}_0$ can be determined numerically. Diagonalizing the Jacobian reveals a real, positive eigenvalue λ_1 for all

σ and β, with the instability timescale $t_*/\tau = -\log(\lambda_1)$ (Fig. 7E). This timescale t_* is of order τ (i.e., destabilization is fast) for $\beta \gtrsim 1$, with weak dependence on σ.

The activity-dependent gain modulation through $Q(t)$ is essential for these pattern-forming dynamics. If we fix $Q = \Gamma^x(0)$, removing its dynamical dependence on the system state, the bump becomes unstable, evidenced by the real, positive eigenvalue in Fig. 7F (orange curve). The system instead converges to a new uniform state $\tilde{m}(\theta) = \tilde{m}'_0$, where $\tilde{m}'_0$ can be determined numerically.

Finally, we note that the overlap order parameter $\tilde{m}(\theta, t)$ determines not only the macroscopic behavior of the system but also the dynamics of the local activities, through

$$x_i(t) = \int_0^{2\pi} d\theta \int_0^{2\pi} d\theta' [\Gamma^\phi]^{-1}(\theta - \theta') x_i^\star(\theta) \tilde{m}(\theta', t). \tag{17}$$

This bidirectional mapping demonstrates how ordered collective computation emerges from, and remains consistent with, the underlying disordered neural responses.

In summary, the main components of the dynamical mean-field theory are

1. A nonlinearity $G_{Q(t)}(h)$ that, unlike classical ring-attractor models, depends on a measure of the network's instantaneous overall activity $Q(t)$ through an emergent gain-control mechanism.
2. Effective weights $J(\Delta\theta)$ supporting Turing pattern formation, with local excitation and longer-range inhibition forming a Mexican-hat profile.
3. A continuous circular symmetry, reflected in the effective weights $J(\Delta\theta)$ depending only on angular differences $\Delta\theta$, that is spontaneously broken, leading to the emergence of bump states.

As described in the introduction, these are the essential ingredients of a *classical* continuous attractor.

5 Grid cells

Our approach for constructing continuous attractors from heterogeneous tuning curves generalizes to symmetries of higher dimension. In fact, we expect this method to extend to general Lie groups. Here we focus on a two-dimensional torus, demonstrating its application to grid cells in the medial entorhinal cortex, which are thought to support a two-dimensional toroidal attractor [22].

Classical continuous-attractor models of grid cells, such as the model of Burak and Fiete [36], involve a two-dimensional sheet of neurons with translation-invariant connectivity. These models generate multiple activity bumps through spatial pattern formation, with two-dimensional shifts of the pattern induced by movement in the environment. Alternative models with a single bump of activity on a twisted torus have also been proposed [42]. These classical models produce idealized, perfectly periodic hexagonal grid patterns. In contrast, experimental observations show that medial-entorhinal cells exhibit more complex, imperfect patterns (Fig. 8A). Grid cells are typically identified by selecting neurons with large "grid scores," a metric quantifying the degree of hexagonal periodicity in spatial firing patterns. In data from various experiments, there often exists a continuous distribution of grid scores with no clear boundary between grid and non-grid responses (Fig. 8B). This contrasts sharply with classical continuous-attractor models, which predict uniformly high grid scores corresponding to a dedicated grid-cell circuit or "module."

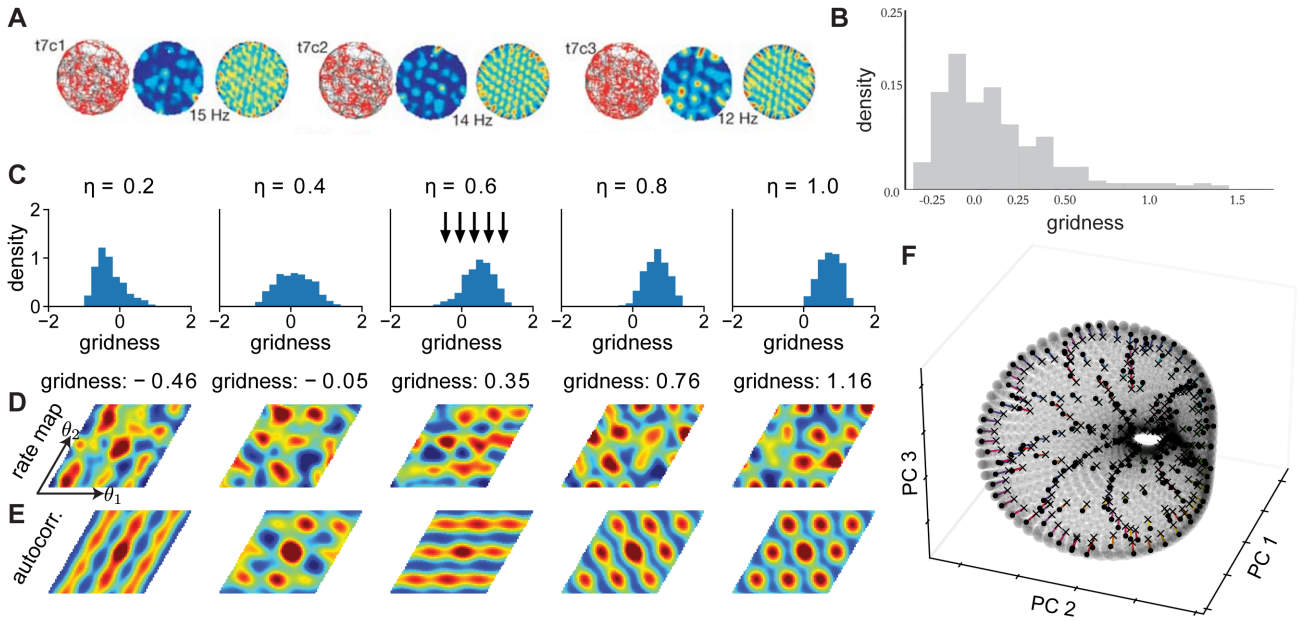

Figure 8: Grid-cell model derived from heterogeneous tuning curves. (**A**) Example grid-cell responses, reproduced from Hafting et al. [16]. Each example shows spiking activity atop the rat's two-dimensional spatial trajectory (left), the firing-rate map (middle), and the spatial autocorrelation (right). (**B**) Distribution of grid scores from experimental recordings, adapted from Nayebi et al. [25], which analyzed data from Mallory et al. [41]. (**C**) Grid-score distributions from the Gaussian-process model for various values of the periodicity parameter η. (**D**) Example firing-rate maps sampled from the Gaussian process with $\eta = 0.6$ (arrow above histogram in C). (**E**) Spatial autocorrelation functions computed from the firing-rate maps in (D), demonstrating enhanced hexagonal structure compared to the raw maps. (**F**) Network dynamics visualized in the space of the first three principal components, showing convergence to the toroidal attractor from various initial conditions (black x's). Colored lines show trajectories over time, with black dots indicating final states.

To generate more realistic grid-cell firing-rate maps, we extended the Gaussian process approach to periodicity in two dimensions. Let the periodic variables be a two-dimensional vector of angles, $\boldsymbol{\theta} = (\theta_1, \theta_2)$. The covariance function of the Gaussian process is given as a periodic summation,

$$\Gamma^x(\Delta\boldsymbol{\theta}) = \sum_{\boldsymbol{n}} A(\boldsymbol{n}) \exp\left(-\sigma^2 \left\| \boldsymbol{M}\left(\Delta\boldsymbol{\theta} - \frac{2\pi}{K}\boldsymbol{n}\right)\right\|_2^2\right), \tag{18}$$

where $\boldsymbol{n} = (n_1, n_2)$ runs over the two-dimensional integer lattice $\mathbb{Z}^2$ and $\boldsymbol{M} = ((1, 1/2), (0, \sqrt{3}/2))$ is a transformation matrix that generates hexagonal symmetry. The weighting function $A(\boldsymbol{n})$ controls the periodicity of the patterns and is given by

$$A(\boldsymbol{n}) = \begin{cases} 1 & \boldsymbol{n} \in (K\mathbb{Z})^2 \\ \eta & \text{otherwise,} \end{cases} \tag{19}$$

where $0 \leq \eta \leq 1$ is the parameter controlling pattern periodicity. When $\eta = 1$, the sampled patterns are periodic with period $2\pi/K$, which is the period of the grid-cell activity pattern. For $\eta < 1$, the firing becomes aperiodic over this period but maintains periodicity over the larger period 2π and retains an underlying toroidal symmetry that is realized in the target manifold and network dynamics in the large-N limit. We set $K = 2$ for generating firing-rate maps and grid-score analyses.

Fig. 8D shows samples of firing-rate maps for various values of η, demonstrating increasing grid-like structure as η approaches unity. Notably, two-dimensional autocorrelograms of these firing fields (Fig. 8E) often display more pronounced hexagonal structure than the firing fields themselves, consistent

with experimental observations [16]. Our approach provides a natural explanation for this phenomenon: the underlying periodic grid pattern is reflected in the autocorrelation function of the process generating the firing-rate maps, rather than in individual maps themselves. A "good grid cell" in our framework is one with a firing-rate map that just happens to resemble the correlation function of the generative process from which it was sampled.

To quantify the grid-like properties of the generated firing-rate maps, we compute grid scores using standard methods (Fig. 8C). A positive grid score indicates hexagonal structure, with higher scores representing more pronounced grid patterns. This approach generates a substantial number of cells that fail to meet the conventional criterion for grid cells (sometimes taken to be grid score > 0), despite being integral components of the toroidal attractor that we develop below. This suggests that the prevalent definition of grid cells may be overly restrictive, discarding neurons that play crucial roles in the emergent, functionally symmetric dynamics.

We now construct attractor networks from samples drawn from the generative process with $K = 1$, which makes the situation more directly analogous to the head-direction model. Using the same least-squares optimization approach described for head-direction cells, but extended to the two-dimensional case, we construct an attractor network based on $N = 25,000$ randomly drawn tuning curves. This larger network size is necessary because the number of neurons required for stable dynamics scales exponentially with the manifold dimension (2 in this case), so we need approximately the square of the neurons required for the head-direction system (Discussion). The resulting network shows rapid convergence to the toroidal manifold from nearby initial conditions (Fig. 8F).

For grid-cell networks constructed using our approach, the connectivity matrix exhibits a six-fold degeneracy in its spectrum at large N, reflecting the hexagonal symmetry of the underlying representation. This is analogous to the doublet degeneracy found in ring networks, but with three-fold multiplicity due to the hexagonal lattice structure. The eigenvectors of this connectivity are random rather than two-dimensional Fourier modes.

When written in Cartesian coordinates $\boldsymbol{x} = \boldsymbol{M}\boldsymbol{\theta}$, the full $N \to \infty$ mean-field description with velocity inputs takes a form similar to classical grid-cell models [36, 42],

$$\partial_t \tilde{m}(\boldsymbol{x}, t) = -\tilde{m}(\boldsymbol{x}, t) + G_{Q(t)} \left(\int d\boldsymbol{x}' \left[1 - \tau \boldsymbol{v}^T(t) \boldsymbol{\nabla}_{\boldsymbol{x}} \right] J(\boldsymbol{x} - \boldsymbol{x}') \tilde{m}(\boldsymbol{x}', t) \right), \tag{20}$$

where $\boldsymbol{v}(t) = (v_x(t), v_y(t))$ is the Cartesian velocity and $\boldsymbol{\nabla}_{\boldsymbol{x}} = (\partial_x, \partial_y)$ is the Cartesian gradient. This dynamical mean-field equation exhibits two-dimensional translation symmetry. Our use of $K = 1$ means that this is a single-bump model; the activity pattern exhibited by $\tilde{m}(\boldsymbol{x}, t)$ is a single bump of activity moving in response to velocity input.

6 Discussion

A common approach in systems neuroscience is to train the weights of recurrent neural networks to perform tasks that animals have mastered, optimizing task performance using standard algorithms. The connectivity and dynamics of the trained network are then compared to experimental data, often identifying structures that either match data or are otherwise interpretable. This method relies on

the inductive biases of the learning algorithm and task constraints to construct solutions that match biological circuits [43, 44]. Our study took an inverse approach [28, 29]. Instead of constraining the network's "macroscopic" task performance, we incorporated the "microscopic" details of single-neuron responses observed in experimental data. This allowed us to investigate the types of network connectivity and dynamics that emerge from, and are consistent with, the experimentally observed heterogeneous neuronal tuning.

Through analysis of both the high-dimensional connectivity and its mean-field description, we demonstrated that despite the disorder in single-neuron tuning, the population can possess dynamics equivalent to a classical continuous attractor, including the presence of continuous symmetry. This symmetry manifests differently than in classical models—while classical models embed their circular symmetry using orderly Fourier patterns across neurons, our model achieves the same symmetry through random, disordered embeddings. This is reflected in the doublet degeneracy of the connectivity matrix's spectrum, providing a robust signature of the underlying circular geometry despite the high level of heterogeneity. The emergent mean-field description, derived in the large-N limit, demonstrates how this hidden symmetry gives rise to the functionally symmetric dynamics characteristic of classical continuous-attractor models.

A key element in this analysis is the use of a statistical generative process to extend mean-field approaches to data-derived models. The overall pipeline we have established—from characterizing single-neuron responses, to developing statistical models of neural heterogeneity, to deriving the necessary circuit mechanisms, to analyzing both the high-dimensional structure and mean-field limit—provides a general framework for understanding how ordered computation emerges within disordered neural circuits. It would be interesting to apply this pipeline to other brain regions and computations. For example, Gaussian-process modeling of neural trajectories has seen some success in the motor system [45].

The heterogeneous attractors derived here functionally mimic classical continuous attractors while representing a much broader class of implementations. The ordered, translation-invariant architectures of classical models, corresponding to Fourier-based embeddings, represent only a special, measure-zero set of solutions within this broader space. This may explain why heterogeneous tuning is prevalent in larger brains such as those of mammals: it represents the generic case rather than a specialized solution. This raises the question: What conditions might favor the emergence of classical, translation-invariant architectures? One possibility is the constraint of limited neuronal resources. This could explain why flies, with their smaller nervous systems, implement an anatomical ring attractor that closely resembles classical models. The restricted number of available neurons may necessitate a more specialized, symmetrical solution to achieve reliable continuous-attractor dynamics [6].

The data-derived connectivity (Eq. 3) exhibits an approximate low-rank structure, where the number of large eigenvalues is much smaller than the number of neurons N. This situates our model within the framework of low-rank recurrent neural networks, a well-studied class of models whose dynamics admit low-dimensional mean-field descriptions [46–49], a property we exploited (Sec. 4.2). Mastrogiuseppe and Ostojic [46] explored how low-rank recurrent neural networks could generate continuous ring manifolds. In a recent preprint, Pezon et al. [50] emphasized that single-neuron representations can

have a higher dimension than the variable they encode (exemplified here by tuning curves having multiple Fourier components, but encoding a one-dimensional variable), and that symmetries in single-neuron representations induce dynamical symmetries (exemplified here by the circular symmetry of the generative process leading to rotation-invariance of the mean-field dynamics, Eq. 14). Our work represents a concrete demonstration of these ideas through the direct optimization of network connectivity to match experimental recordings. Finally, while previous work has proposed methods for fitting low-rank networks to neural data through machine learning [48, 51], our approach provides an explicit analytical solution for connectivity (Eq. 3) that, when combined with the generative process, leads to continuous-attractor dynamics.

While our optimization-based approach determines the synaptic weights required for continuous-attractor dynamics that reproduce neuronal responses (Eq. 3), understanding how these weights could be established through local, biologically plausible plasticity mechanisms remains an open question. The Hebbian-like structure of the solution (Eq. 3) suggests that some form of correlation-based learning might suffice.

As demonstrated in our grid cell analysis, the number of neurons required for stable continuous-attractor dynamics scales exponentially with the intrinsic dimension d of the target manifold. This scaling poses a significant challenge for modeling higher-dimensional representations ($d > 2$) within biological constraints. However, this exponential scaling assumes that the Gaussian-process covariance function allows for arbitrary conjunctive tuning across all manifold dimensions. The scaling can be dramatically reduced through structured constraints on the form of neuronal tuning. In the extreme case, if a d-dimensional manifold can be factorized into d independent one-dimensional manifolds, the curse of dimensionality is entirely circumvented. More realistically, covariance functions can be constructed that permit partial but not complete mixing across dimensions. Such constrained mixing would predict a population where some neurons exhibit specialized responses to individual variables while most show some degree of conjunctive tuning. This prediction aligns remarkably well with observations in regions like the medial entorhinal cortex, where grid cells coexist with neurons displaying diverse conjunctive responses to spatial and non-spatial variables [23, 25]. Extending our theoretical framework to study mixed selectivity represents a promising direction for future research.

Acknowledgments

We thank A. Litwin-Kumar, J. Zavatone-Veth, J. Cotler and H. Grier for interesting conversations. This research was supported by the Gatsby Charitable Foundation (D.G.C., L.F.A. and H.S.) and the Kavli Foundation (D.G.C. and L.F.A.). H.S. was additionally supported by the Swartz Foundation and the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University.

References

[1] Mikail Khona and Ila R Fiete. Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*, 23(12):744–766, 2022.

[2] 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.

[3] Xiaohui Xie, Richard HR Hahnloser, and H Sebastian Seung. Double-ring network model of the head-direction system. *Physical Review E*, 66(4):041902, 2002.

[4] Samuel A Ocko, Kiah Hardcastle, Lisa M Giocomo, and Surya Ganguli. Emergent elasticity in the neural code for space. *Proceedings of the National Academy of Sciences*, 115(50):E11798–E11806, 2018.

[5] Raymond Wang and Louis Kang. Multiple bumps can enhance robustness to noise in continuous attractor networks. *PLOS Computational Biology*, 18(10):e1010547, 2022.

[6] Marcella Noorman, Brad K Hulse, Vivek Jayaraman, Sandro Romani, and Ann M Hermundstad. Maintaining and updating accurate internal representations of continuous variables with a handful of neurons. *Nature Neuroscience*, pages 1–11, 2024.

[7] Kechen Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *Journal of Neuroscience*, 16(6):2112–2126, 1996.

[8] Omri Barak and Sandro Romani. Mapping low-dimensional dynamics to high-dimensional neural activity: A derivation of the ring model from the neural engineering framework. *Neural Computation*, 33(3):827–852, 2021.

[9] Johannes D Seelig and Vivek Jayaraman. Neural dynamics for landmark orientation and angular path integration. *Nature*, 521(7551):186–191, 2015.

[10] Jonathan Green, Adsuko Adachi, Kunal K Shah, Jonathan D Hikokawa, Pablo S Magani, and Gaby Maimon. A neural circuit architecture for angular integration in drosophila. *Nature*, 546: 101–106, 2017.

[11] Daniel Turner-Evans, Stephanie Wegener, Herve Roualult, Romain Franconville, Tanya Wolff, Johannes D. Seelig, Shaul Druckmann, and VIvek Jayaraman. Angular velocity integration in a fly heading circuit. *eLife*, 23495, 2017.

[12] Luigi Petrucco, Hagar Lavian, You Kure Wu, Fabian Svara, Vilim Štih, and Ruben Portugues. Neural dynamics and architecture of the heading direction circuit in zebrafish. *Nature neuroscience*, 26(5):765–773, 2023.

[13] James B Ranck. Head direction cells in the deep cell layer of dorsolateral pre-subiculum in freely moving rats. *Electrical activity of the archicortex*, 1985.

[14] Jeffrey S Taube, Robert U Muller, and James B Ranck. Head-direction cells recorded from the postsubiculum in freely moving rats. i. description and quantitative analysis. *Journal of Neuroscience*, 10(2):420–435, 1990.

[15] Jeffrey S Taube, Robert U Muller, and James B Ranck. Head-direction cells recorded from the postsubiculum in freely moving rats. ii. effects of environmental manipulations. *Journal of Neuroscience*, 10(2):436–447, 1990.

[16] Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005.

[17] Edvard I Moser, Emilio Kropff, and May-Britt Moser. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, 31(1):69–89, 2008.

[18] Rishidev Chaudhuri, Berk Gerçek, Biraj Pandey, Adrien Peyrache, and Ila Fiete. The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nature neuroscience*, 22(9):1512–1520, 2019.

[19] Adrian J Duszkiewicz, Pierre Orhan, Sofia Skromne Carrasco, Eleanor H Brown, Eliott Owczarek, Gilberto R Vite, Emma R Wood, and Adrien Peyrache. Local origin of excitatory–inhibitory tuning equivalence in a cortical network. *Nature Neuroscience*, 27(4):782–792, 2024.

[20] KiJung Yoon, Michael A Buice, Caswell Barry, Robin Hayman, Neil Burgess, and Ila R Fiete. Specific evidence of low-dimensional continuous attractor dynamics in grid cells. *Nature neuroscience*, 16(8):1077–1084, 2013.

[21] Richard J Gardner, Li Lu, Tanja Wernle, May-Britt Moser, and Edvard I Moser. Correlation structure of grid cells is preserved during sleep. *Nature neuroscience*, 22(4):598–608, 2019.

[22] Richard J Gardner, Erik Hermansen, Marius Pachitariu, Yoram Burak, Nils A Baas, Benjamin A Dunn, May-Britt Moser, and Edvard I Moser. Toroidal topology of population activity in grid cells. *Nature*, 602(7895):123–128, 2022.

[23] Kiah Hardcastle, Niru Maheswaranathan, Surya Ganguli, and Lisa M Giocomo. A multiplexed, heterogeneous, and adaptive code for navigation in medial entorhinal cortex. *Neuron*, 94(2): 375–387, 2017.

[24] Yi Gu, Sam Lewallen, Amina A Kinkhabwala, Cristina Domnisoru, Kijung Yoon, Jeffrey L Gauthier, Ila R Fiete, and David W Tank. A map-like micro-organization of grid cells in the medial entorhinal cortex. *Cell*, 175(3):736–750, 2018.

[25] Aran Nayebi, Alexander Attinger, Malcolm Campbell, Kiah Hardcastle, Isabel Low, Caitlin S Mallory, Gabriel Mel, Ben Sorscher, Alex H Williams, Surya Ganguli, et al. Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks. *Advances in Neural Information Processing Systems*, 34:12167–12179, 2021.

[26] Nischal Mainali, Rava Azeredo da Silveira, and Yoram Burak. Universal statistics of hippocampal place fields across species and dimensionalities. *bioRxiv*, pages 2024–06, 2024.

[27] Ran Darshan and Alexander Rivkind. Learning to represent continuous variables in heterogeneous neural networks. *Cell Reports*, 110612, 2022.

[28] Dimitry Fisher, Itsaso Olasagasti, David W Tank, Emre RF Aksay, and Mark S Goldman. A modeling framework for deriving the structural and functional architecture of a short-term memory microcircuit. *Neuron*, 79(5):987–1000, 2013.

[29] K Rajan, CD Harvey, and DW Tank. Recurrent network models of sequence generation and memory. *Neuron*, 90(1):128–412, 2016.

[30] Adrian Duszkiewicz, Sofia Skromne Carrasco, and Adrien Peyrache. Large-scale recordings of head-direction cells in mouse postsubiculum, 2024. URL https://doi.org/10.48324/dandi.000939/0.240528.1542.

[31] Jeffrey S Taube. The head direction signal: origins and sensory-motor integration. *Annu. Rev. Neurosci.*, 30(1):181–207, 2007.

[32] Jonathan Kadmon and Haim Sompolinsky. Transition to chaos in random neuronal networks. *Physical Review X*, 5(4):041030, 2015.

[33] Aaron S Andalman, Vanessa M Burns, Matthew Lovett-Barron, Michael Broxton, Ben Poole, Samuel J Yang, Logan Grosenick, Talia N Lerner, Ritchie Chen, Tyler Benster, et al. Neuronal dynamics regulating brain and behavioral state transitions. *Cell*, 177(4):970–985, 2019.

[34] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[35] John J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092, 1984.

[36] Yoram Burak and Ila R Fiete. Accurate path integration in continuous attractor network models of grid cells. *PLoS computational biology*, 5(2):e1000291, 2009.

[37] Ben Sorscher, Gabriel C Mel, Samuel A Ocko, Lisa M Giocomo, and Surya Ganguli. A unified theory for the computational and mechanistic origins of grid cells. *Neuron*, 111(1):121–137, 2023.

[38] Adityanarayanan Radhakrishnan, Mikhail Belkin, and Caroline Uhler. Overparameterized neural networks implement associative memory. *Proceedings of the National Academy of Sciences*, 117 (44):27162–27170, 2020.

[39] William Skaggs, Bruce Mcnaughton, and Katalin Gothard. An information-theoretic approach to deciphering the hippocampal code. *Advances in neural information processing systems*, 5, 1992.

[40] Larry F Abbott, Davi D Bock, Edward M Callaway, Winfried Denk, Catherine Dulac, Adrienne L Fairhall, Ila Fiete, Kristen M Harris, Moritz Helmstaedter, Viren Jain, et al. The mind of a mouse. *Cell*, 182(6):1372–1376, 2020.

[41] Caitlin S Mallory, Kiah Hardcastle, Malcolm G Campbell, Alexander Attinger, Isabel IC Low, Jennifer L Raymond, and Lisa M Giocomo. Mouse entorhinal cortex encodes a diverse repertoire of self-motion signals. *Nature communications*, 12(1):671, 2021.

[42] Alexis Guanella, Daniel Kiper, and Paul Verschure. A model of grid cells based on a twisted torus topology. *International journal of neural systems*, 17(04):231–240, 2007.

[43] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

[44] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

[45] Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *BioRxiv*, page 214262, 2017.

[46] Francesca Mastrogiuseppe and Srdjan Ostojic. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, 2018.

[47] Manuel Beiran, Alexis Dubreuil, Adrian Valente, Francesca Mastrogiuseppe, and Srdjan Ostojic. Shaping dynamics with multiple populations in low-rank recurrent networks. *Neural Computation*, 33(6):1572–1615, 2021.

[48] Alexis Dubreuil, Adrian Valente, Manuel Beiran, Francesca Mastrogiuseppe, and Srdjan Ostojic. The role of population structure in computations through neural dynamics. *Nature neuroscience*, 25(6):783–794, 2022.

[49] David G Clark and Manuel Beiran. Structure of activity in multiregion recurrent neural networks. *arXiv preprint arXiv:2402.12188*, 2024.

[50] Louis Pezon, Valentin Schmutz, and Wulfram Gerstner. Linking neural manifolds to circuit structure in recurrent networks. *bioRxiv*, pages 2024–02, 2024.

[51] Adrian Valente, Jonathan W Pillow, and Srdjan Ostojic. Extracting computational mechanisms from neural data using low-rank rnns. *Advances in Neural Information Processing Systems*, 35: 24072–24086, 2022.

Supplementary Information

SI.1 Data processing and analysis

We analyzed neural recordings collected by Duszkiewicz et al. [19] (see [30] for the publicly available dataset) from mice exploring a square environment, comprising 1,533 excitatory neurons that were previously identified as head-direction cells. Nine of these mice had viral expression for optogenetic manipulations, though we excluded periods of optogenetic stimulation and processed this data identically to non-optogenetic sessions.

Neural and behavioral data were stored in Neurodata Without Borders (NWB) format. For each recording session, we extracted the animal's head direction and spike times during periods of square environment exploration. Head direction angles were binned into 100 equally sized bins spanning the full 2π radians. For each neuron, we computed raw tuning curves by dividing the spike count in each angular bin by the time spent in that bin (occupancy).

To address sampling noise while preserving genuine high-frequency features in the tuning curves, we implemented a cross-validated smoothing procedure. For each neuron, we:

1. Divided the recording session into eight temporal segments of equal duration, labeled alternately as partition 1 or 2 (i.e., 1-2-1-2-1-2-1-2)
2. Computed raw tuning curves independently for each partition
3. Applied Gaussian smoothing kernels of varying widths to the tuning curve from each partition, with widths logarithmically spaced from 0.1 to 100 bins (corresponding to $2\pi/1000$ to 2π radians)
4. Evaluated the Poisson log-likelihood of the spiking data from one partition given the smoothed tuning curve from the other partition
5. Selected the smoothing width that maximized this cross-validated log-likelihood
6. Computed final tuning curves by averaging the optimally smoothed curves from both partitions

Finally, we normalized all tuning curves to have a mean firing rate of unity (arbitrary units) across all head directions. This normalization allowed direct comparison of tuning properties across neurons and animals. These normalized, smoothly varying tuning curves formed the foundation for all subsequent analyses in this study.

SI.2 Data-derived network construction

For the nonlinearity in this section, we used a "softplus" function

$$\phi(x) = \beta^{-1} \log(1 + \exp(\beta x)), \tag{21}$$

with $\beta = 2$. This is similar to the nonlinearity used in the generative process (Sec. 3.2) but without a bias term. Since the tuning curves $\phi^\star(\theta)$ are measured experimentally rather than sampled, and have

already been normalized to unity when averaged over θ during data processing (SI.1), the normalization functional described in Sec. 3.2 is not required here.

We performed an additional normalization step for the data-derived network construction. As demonstrated in Sec. 3.1, the data are consistent with distributional circular symmetry, and since N_{data} is large, we expect $N_{\text{data}}^{-1} \sum_{i=1}^{N_{\text{data}}} \phi_i^\star(\theta) \approx 1$ for all θ. However, while the integral over θ is exactly unity by construction, the population average at each θ exhibits small fluctuations around unity. To enforce exact population averaging, we simultaneously normalized both the neuronal (i) and angular (θ) averages to unity using the Sinkhorn-Knopp algorithm, which iteratively normalizes these averages until convergence. This results in a tiny correction to each tuning curve.

SI.3 Computing the weight matrix

The solution to the constrained optimization problem for the weight matrix of the data-derived network can be computed using several equivalent methods. Using bold symbols to denote matrices, we define $\boldsymbol{\Phi}$ ($N_\theta \times N$) as the firing rate matrix whose columns contain the tuning curves $\phi_i^\star(\theta)$, and $\boldsymbol{X}$ ($N_\theta \times N$) as the input-current matrix containing the corresponding input currents $x_i^\star(\theta)$, where $N = N_{\text{data}}$.

For $N_\theta > N$, we first construct and invert

$$\boldsymbol{C} = \frac{1}{N_\theta} \boldsymbol{\Phi}^T \boldsymbol{\Phi} + \lambda \boldsymbol{I}_N, \tag{22}$$

where $\boldsymbol{C}$ is an $N \times N$ matrix. The unconstrained solution is then

$$\boldsymbol{J}^{\text{unconstrained}} = \frac{1}{N_\theta} \boldsymbol{X}^T \boldsymbol{\Phi} \boldsymbol{C}^{-1} \tag{23}$$

For $N_\theta < N$, we instead construct and invert

$$\boldsymbol{K} = \frac{1}{N} \boldsymbol{\Phi} \boldsymbol{\Phi}^T + \frac{\lambda N_\theta}{N} \boldsymbol{I}_{N_\theta}, \tag{24}$$

where $\boldsymbol{K}$ is an $N_\theta \times N_\theta$ matrix and $\boldsymbol{I}_{N_\theta}$ is the $N_\theta \times N_\theta$ identity matrix. The unconstrained solution is

$$\boldsymbol{J}^{\text{unconstrained}} = \frac{1}{N} \boldsymbol{X}^T \boldsymbol{K}^{-1} \boldsymbol{\Phi}, \tag{25}$$

where $\boldsymbol{J}^{\text{unconstrained}}$ is an $N \times N$ matrix. We then compute (without requiring explicit $N \times N$ inversion)

$$\boldsymbol{C}^{-1} = \boldsymbol{I}_N - \frac{1}{N} \boldsymbol{\Phi}^T \boldsymbol{K}^{-1} \boldsymbol{\Phi}, \tag{26}$$

where $\boldsymbol{C}$ is an $N \times N$ matrix.

In both cases, the final constrained solution is

$$J_{ij} = J_{ij}^{\text{unconstrained}} - \frac{C_{ij}^{-1} J_{ii}^{\text{unconstrained}}}{C_{ii}^{-1}}. \tag{27}$$

SI.4 Variants on the weight matrix

For the circulant and noisy circulant variants, we multiply by a factor of 1.275 to ensure bump formation. This factor was chosen since it leads to bump formation and produces bumps of similar height to those in networks with the data-derived weight matrix.

SI.5 Derivation of dynamical mean-field theory equations

We begin by deriving the dynamical mean-field equations for zero velocity input ($\omega(t) = 0$), which capture the essential features of the derivation while simplifying the presentation. This allows us to describe the transient evolution of the overlap order parameter from arbitrary initial conditions. We then extend the analysis to incorporate velocity inputs.

Starting with the network dynamics (Eq. 1) and the solution for the weights J_{ij} (Eq. 3), we have

$$(1+\tau\partial_t)x_i(t) = \sum_{j=1}^{N} \underbrace{\left[\frac{1}{N}\int_0^{2\pi} d\theta \int_0^{2\pi} d\theta' K^{-1}_{i,\lambda}(\theta,\theta')x^\star_i(\theta)\phi^\star_j(\theta')\right]}_{J_{ij}} \phi_j(t). \tag{28}$$

In the limit $N \to \infty$ with $\lambda = 0$, the inverse kernel $K^{-1}_{i,\lambda}(\theta,\theta')$ converges to $[\Gamma^\phi]^{-1}(\theta-\theta')$, where $\Gamma^\phi(\Delta\theta)$ is invertible due to all Fourier components being nonzero. Moving the sum over j inside the integrals and using the definition of $m(\theta,t)$ from Eq. 6, we obtain

$$(1+\tau\partial_t)x_i(t) = \int_0^{2\pi} d\theta \int_0^{2\pi} d\theta' [\Gamma^\phi]^{-1}(\theta-\theta')x^\star_i(\theta)m(\theta',t). \tag{29}$$

We introduce a temporally smoothed version of the overlap order parameter,

$$\tilde{m}(\theta,t) = \int_0^t \frac{dt'}{\tau} e^{-(t-t')/\tau} m(\theta,t) + e^{-t/\tau} m(\theta,0). \tag{30}$$

This satisfies

$$(1+\tau\partial_t)\tilde{m}(\theta,t) = m(\theta,t), \tag{31}$$

as stated in the main text. Applying $(1+\tau\partial_t)^{-1}$ to both sides of the dynamics equation (Eq. 29) yields

$$x_i(t) = \int_0^{2\pi} d\theta \int_0^{2\pi} d\theta' [\Gamma^\phi]^{-1}(\theta-\theta')x^\star_i(\theta)\tilde{m}(\theta',t). \tag{32}$$

Applying $\phi(\cdot)$ to both sides of this equation, and then taking the inner products of both sides with $\phi^\star_i(\theta)$ and dividing them by N, gives

$$m(\theta,t) = \frac{1}{N}\sum_{i=1}^{N} \phi(x^\star_i(\theta))\phi\left(\int_0^{2\pi} du \int_0^{2\pi} du' [\Gamma^\phi]^{-1}(u-u')x^\star_i(u)\tilde{m}(u',t)\right). \tag{33}$$

The key step in deriving the mean-field equations is to replace the empirical average over neurons (i index) with an expectation over the generating Gaussian process. That is,

$$m(\theta,t) = \left\langle \phi(\underbrace{x^\star(\theta)}_{\text{term 1}})\phi\left(\underbrace{\int_0^{2\pi} du \int_0^{2\pi} du' [\Gamma^\phi]^{-1}(u-u')x^\star(u)\tilde{m}(u',t)}_{\text{term 2}}\right)\right\rangle_{x^\star\sim\mathcal{GP}(0,\Gamma^x)}. \tag{34}$$

This replacement is exact for $N \to \infty$. Terms 1 and 2 are jointly Gaussian with zero mean. Their

variances and covariance fully determine the right-hand side of the equation and are given by

$$\left\langle [\text{term 1}]^2 \right\rangle = \Gamma^x(0), \tag{35}$$

$$\left\langle [\text{term 2}]^2 \right\rangle = \int_0^{2\pi} du \int_0^{2\pi} du' \int_0^{2\pi} dv \int_0^{2\pi} dv' [\Gamma^\phi]^{-1}(u-u')[\Gamma^\phi]^{-1}(v-v')\Gamma^x(u-v)\tilde{m}(u',t)\tilde{m}(v',t), \tag{36}$$

$$\langle [\text{term 1}][\text{term 2}] \rangle = \int_0^{2\pi} du \int_0^{2\pi} du' [\Gamma^\phi]^{-1}(u-u')\Gamma^x(\theta-u)\tilde{m}(u',t). \tag{37}$$

We define kernels in Fourier space:

$$U_n = \frac{\Gamma^x_n}{(\Gamma^\phi_n)^2}, \tag{38}$$

$$J_n = \frac{\Gamma^x_n}{\Gamma^\phi_n}. \tag{39}$$

The second equation (Eq. 39) defines the effective rotation-invariant weights stated in the main text (Eq. 13). These kernels allow us to express the correlations between terms 1 and 2 as

$$\left\langle [\text{term 2}]^2 \right\rangle = \int_0^{2\pi} d\theta \int_0^{2\pi} d\theta' U(\theta-\theta')\tilde{m}(\theta,t)\tilde{m}(\theta',t) \equiv Q(t), \tag{40}$$

$$\langle [\text{term 1}][\text{term 2}] \rangle = \int_0^{2\pi} d\theta' J(\theta-\theta')\tilde{m}(\theta',t). \tag{41}$$

Now, let $\mathcal{G}(C_{11}, C_{22}, C_{12})$ denote the expectation

$$\mathcal{G}(C_{11}, C_{22}, C_{12}) = \langle \phi(z_1)\phi(z_2) \rangle_{\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{pmatrix} \right)}. \tag{42}$$

This can be explicitly evaluated using the nonlinearity $\phi(x) = 1 + \text{erf}(\beta\sqrt{\pi}x/2)$, which yields

$$\mathcal{G}(C_{11}, C_{22}, C_{12}) = 1 + \frac{2}{\pi}\tan^{-1}\left(\frac{\beta^2 C_{12}}{\sqrt{(\beta^2 C_{11} + \frac{2}{\pi})(\beta^2 C_{22} + \frac{2}{\pi}) - (\beta^2 C_{12})^2}} \right). \tag{43}$$

Then, Eq. 34 can be written as

$$m(\theta,t) = \mathcal{G}\left(\Gamma^x(0), Q(t), \int_0^{2\pi} d\theta' J(\theta-\theta')\tilde{m}(\theta',t) \right). \tag{44}$$

Defining

$$G_{Q(t)}(h) = \mathcal{G}(\Gamma^x(0), Q(t), h) \tag{45}$$

and using the relationship between $m(\theta,t)$ and $\tilde{m}(\theta,t)$, we obtain

$$(1+\tau\partial_t)\tilde{m}(\theta,t) = G_{Q(t)}\left(\int_0^{2\pi} d\theta' J(\theta-\theta')\tilde{m}(\theta',t) \right). \tag{46}$$

This is the mean-field dynamics from the main text (Eq. 14) with $\omega(t) = 0$.

SI.5.1 Velocity-dependent dynamics

We now extend this derivation to include nonzero angular velocity $\omega(t)$. The network dynamics are modified to include an angular velocity-dependent term,

$$(1+\tau\partial_t)x_i(t) = \int_0^{2\pi} d\theta \int_0^{2\pi} d\theta' \Gamma_\phi^{-1}(\theta-\theta')(1+\tau\omega(t)\partial_\theta)x_i^\star(\theta)m(\theta',t). \tag{47}$$

To solve Eq. 47, we again apply $(1+\tau\partial_t)^{-1}$ to both sides, and now introduce a variable $\tilde{M}(\theta,t)$ defined as

$$\tilde{M}(\theta,t) = \int_0^t \frac{dt'}{\tau} e^{-(t-t')/\tau}\tau\omega(t')m(\theta,t') + e^{-t/\tau}\tau\omega(0)m(\theta,0), \tag{48}$$

which satisfies

$$(1+\tau\partial_t)\tilde{M}(\theta,t) = \tau\omega(t)m(\theta,t). \tag{49}$$

Then, we have

$$x_i(t) = \int_0^{2\pi} d\theta \int_0^{2\pi} d\theta' [\Gamma^\phi]^{-1}(\theta-\theta')\left(x_i^\star(\theta)\tilde{m}(\theta',t) + \partial_\theta x_i^\star(\theta)\tilde{M}(\theta',t)\right). \tag{50}$$

We now massage Eq. 48 by expanding $\omega(t')$ in a Taylor series around t (dropping the initial transient for simplicity),

$$\begin{aligned}\tilde{M}(\theta,t) &= \left[\int_0^t \frac{dt'}{\tau} e^{-(t-t')/\tau}\tau\left[\omega(t) + (\partial_t\omega(t))(t-t') + \cdots\right] m(\theta,t')\right] \\ &= \left[\tau\omega(t)\tilde{m}(t,\theta) + \tau^2(\partial_t\omega(t))\int_0^t \frac{ds}{\tau}\frac{s}{\tau}e^{-s/\tau}m(\theta,t-s) + \cdots\right].\end{aligned} \tag{51}$$

The expansion in square brackets is in powers $(\tau\partial_t)^n\tau\omega(t)$ for $n = 0,1,2,\ldots$, which we will assume can be truncated after $n=0$. Note that this assumption means that the angular acceleration, given by $\partial_t\omega(t)$ (times appropriate factors of τ), is small, but the angular velocity itself may be large. This gives

$$\tilde{M}(\theta,t) \approx \tau\omega(t)\tilde{m}(t,\theta). \tag{52}$$

Substituting Eq. 52 back into Eq. 50, we obtain

$$x_i(t) = \int_0^{2\pi} d\theta \int_0^{2\pi} d\theta' [\Gamma^\phi]^{-1}(\theta-\theta')\left(1+\tau\omega(t)\partial_\theta\right)x_i^\star(\theta)\tilde{m}(\theta',t). \tag{53}$$

Following the same averaging procedure as in the zero-velocity case, but now accounting for the velocity-dependent terms in Eq. 53, the correlations are modified in two ways:

1. The kernel $U(\Delta\theta)$ that determines the activity-dependent gain parameter $Q(t)$ (the variance of input currents) is modified to include an angular velocity-dependent term,

$$U_{\omega(t)}(\Delta\theta) = \left(1-\tau^2\omega(t)^2\partial_{\Delta\theta}^2\right)U(\Delta\theta), \tag{54}$$

   where $U(\Delta\theta)$ is the original kernel defined through its Fourier transform $U(n) = \Gamma_x(n)/\Gamma_\phi^2(n)$.

2. The correlation between input currents and target tuning curves acquires an additional angular velocity-dependent term,

$$\int_0^{2\pi} d\theta \int_0^{2\pi} d\theta'(1-\tau\omega(t)\partial_\theta)J(\theta-\theta')\tilde{m}(\theta',t). \tag{55}$$

These modifications lead to the full dynamical mean-field equation for the overlap order parameter,

$$(1+\tau\partial_t)\tilde{m}(\theta,t) = G_{Q(t)}\left(\int_0^{2\pi} d\theta' \left[1-\omega(t)\partial_\theta\right] J(\theta-\theta')\tilde{m}(\theta',t)\right), \tag{56}$$

where $Q(t)$ uses $U_{\omega(t)}(\Delta\theta)$ (Eq. 54) in the quadratic term.

SI.6 Jacobian analysis of the mean-field dynamics

To analyze the stability of solutions to the mean-field equations, we need to compute the Jacobian matrix $M(\theta,\theta')$, which describes how perturbations to the overlap order parameter evolve in time. This requires expressions for derivatives of the effective nonlinearity $G_{Q(t)}(h)$ with respect to both h and $Q(t)$. Recall that $G_{Q(t)}(h)$ is defined through the function $\mathcal{G}(C_{11},C_{22},C_{12})$, which computes the expectation in Eq. 42.

We define auxiliary quantities

$$A = \left(\beta^2 C_{11} + \frac{2}{\pi}\right)\left(\beta^2 C_{22} + \frac{2}{\pi}\right) \text{ and } B = A - \left(\beta^2 C_{12}\right)^2. \tag{57}$$

Then, the required derivatives are

$$\partial_{C_{12}}\mathcal{G}(C_{11},C_{22},C_{12}) = \langle \phi'(z_1)\phi'(z_2)\rangle_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{C})} = \frac{2\beta^2}{\pi}\frac{1}{\sqrt{B}}, \tag{58}$$

$$\partial_{C_{11}}\mathcal{G}(C_{11},C_{22},C_{12}) = \frac{1}{2}\langle \phi''(z_1)\phi(z_2)\rangle_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{C})} = -\frac{\beta^4}{\pi}\frac{C_{12}}{A\sqrt{B}}\left(\beta^2 C_{22} + \frac{2}{\pi}\right), \tag{59}$$

where $\boldsymbol{C} = ((C_{11},C_{12}),(C_{12},C_{22}))$.

For the bump solution $\tilde{m}(\theta) = \Gamma^\phi(\theta)$, the Jacobian takes the form

$$M(\theta,\theta') = -\delta(\theta-\theta') + \partial_{C_{12}}\mathcal{G}(\Gamma^x(0),\Gamma^x(0),\Gamma^x(\theta))J(\theta-\theta') + 2\partial_{C_{11}}\mathcal{G}(\Gamma^x(0),\Gamma^x(0),\Gamma^x(\theta))J(\theta'). \tag{60}$$

Each of the three terms represents a distinct contribution to the dynamics: exponential decay (first term), direct coupling through the effective weights (second term), and modulation through the activity-dependent gain $Q(t)$ (third term).

For the homogeneous solution $\tilde{m}(\theta) = \tilde{m}_0$, we first determine $\tilde{m}_0$ by numerically solving the self-consistent equations

$$\tilde{m}_0 = \mathcal{G}(Q_0,\Gamma^x(0),h_0), \tag{61}$$

$$Q_0 = \tilde{m}_0^2\hat{U}_0, \tag{62}$$

$$h_0 = \tilde{m}_0\hat{J}_0. \tag{63}$$

The Jacobian at this solution is then

$$M(\theta,\theta') = -\delta(\theta-\theta') + \partial_{C_{12}}\mathcal{G}(Q_0,\Gamma^x(0),h_0)J(\theta-\theta') + 2\partial_{C_{11}}\mathcal{G}(Q_0,\Gamma^x(0),h_0)\hat{U}_0\tilde{m}_0. \tag{64}$$

Finally, we explain how to analyze the dynamical mean-field theory equations in scenarios when the dynamics of $Q(t)$ are excluded by fixing $Q = \Gamma^x(0)$, as discussed in the main text. We simply omit the

terms involving $\partial_{C_{11}}\mathcal{G}(C_{11}, C_{22}, C_{12})$ in both Jacobians (and, if we would like to find the new stable uniform state, denoted by $\tilde{m}_0'$ in the main text, we fix $Q_0 = \Gamma^x(0)$ in the self-consistent equations above rather than allowing it to depend on $\tilde{m}_0$). The presence of a positive eigenvalue of the Jacobian for the bump state when we fix $Q = \Gamma^x(0)$, shown in Fig. 7F, demonstrates that the activity-dependent gain modulation through $Q(t)$ is essential for bump stability.