Linking Neural Manifolds to Circuit Structure in Recurrent Networks

Louis Pezon[1*], Valentin Schmutz[1,2†] and Wulfram Gerstner[1†]

[1*]School of Computer and Communication Sciences and School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland.
[2]UCL Queen Square Institute of Neurology, University College London WC1E 6BT London, United Kingdom.

*Corresponding author(s). E-mail(s): louis.pezon@epfl.ch;
Contributing authors: v.schmutz@ucl.ac.uk; wulfram.gerstner@epfl.ch;
†These authors contributed equally to this work.

Abstract

While analyses of large-scale neural recording indicate that the activity of heterogeneous populations of neurons follows collective dynamics on low-dimensional neural manifolds, it has remained challenging to reconcile this picture with the classical view of precisely tuned neurons interacting with each other in an ordered circuit structure. Using a modelling approach, we connect these two contrasting views. First, we propose a theoretical framework that explicitly links the circuit structure and the emergent low-dimensional dynamics of the population activity in models of recurrent neural networks. The theory predicts a non-unique relationship between the two, which we illustrate with concrete examples. We then propose a method for retrieving the circuit structure from recordings of the population activity and test it on artificial data. Our approach provides not only a unifying framework for circuit and field models on one side, and low-rank networks on the other side, but also opens the perspective to identify principles of circuit structure from large-scale recordings.

Keywords: neural manifold, recurrent neural networks, low-rank connectivity, neural field model, mean-field theory

Introduction

Our understanding of the link between brain activity and behaviour is continuously shaped by the progress of experimental methods. While, for a long time, neuroscientists could only record one or a few neurons at the same time, the recent availability of tools such as high-density probes and two-photon calcium imaging has enabled the simultaneous recording of thousands of neurons [1]. As a consequence, systems neuroscience currently offers two complementary perspectives on the structure of neuronal activity [2, 3].

The first perspective focuses on the representation and processing of information by single neurons characterised by individual functional properties. As a starting point, landmark studies in neuroscience have established that the responses of single neurons are modulated by features of sensory stimuli [4–6], spatial position and orientation in the environment [7–11], or higher-level task-related variables [12–14]. The dependence of single-neuron responses to stimulus- or task-related features, usually referred to as *tuning*, can generally be summarised by a few properties, such as the centre of their receptive field. For example, in flies and mammals, the sense of spatial orientation arises from a population of head-direction cells, where each cell responds preferentially to a specific head direction (HD) [9, 10, 15]. Cells can thus be sorted according to their functional role in the head direction circuit by placing them on a ring (Fig.1A), where neighbouring cells have similar preferred HD [16]. Another example is given by neurons in the mammalian primary visual cortex (V1), whose receptive field can be characterised by a few features; e.g., location in the visual space, frequency and orientation of the preferred stimulus [17]. Each cell can thus be embedded in an abstract four-dimensional space according to these properties (Fig.1B). Single-neuron responses might also be modulated by some task- or stimulus-independent variables, that are not controlled by the experimenter – such as internal brain state [18–21]. Therefore, single-neuron functional properties can characterise other response features besides stimulus tuning. In this view, each neuron can be assigned to a "location", i.e. be embedded, in an abstract space where neighbourhood reflects the functional similarity between neurons. We call it the *functional similarity space* or, more simply, *similarity space*. Importantly, the embedding of the neurons may not correspond to their physical location in the brain; for instance in mammalian V1, preferred orientations are distributed across the cortical sheet, forming pinwheel-like patterns [22, 23].

The second perspective takes a descriptive, population-level approach, pushed forward by the advent of large-scale recordings [24]. In this perspective, the joint activity of the recorded neurons, or *population activity*, is represented at any given time as a point in a high-dimensional state-space. Several studies revealed that the activity of heterogeneous, large populations of neurons is often restricted to a low-dimensional space, often referred to as the *neural manifold*, embedded within the state-space of the population activity [25–28]. As a point in the low-dimensional neural manifold, the activity can be parameterised by a set of time-dependent latent variables, which encode computationally relevant quantities at the population level; e.g., the animal's head direction (Fig.1D). These latent variables can be read out from the population activity (using either a linear or non-linear decoder), revealing the low-dimensional dynamics driving their trajectory (Fig.1E) [21, 29–31]. Hence, in this view, neural computation is described by a dynamical flow in the low-dimensional neural manifold (Fig.1F), emerging from the collective dynamics of a large number of neurons [31–35].

To summarise, the first perspective focuses on how the functional properties of individual neurons shape neuronal activity; whereas the second perspective focuses on a population-level

description of neural representations and computations [2]. The concept of tuning potentially provide a direct link between the two perspectives [36]. A tuning curve traditionally describes how a single neuron's response depends on a particular stimulus variable. Thus, the set of tuning curves of all the neurons in a population directly maps each possible stimulus to the corresponding population activity. If the set of all possible stimuli is parameterised, say, by a one-dimensional circular variable (for example, the head direction in the case of HD cells), then the set of all the possible population activities will in turn form a ring (Fig.1D). Yet, such a static transformation from stimulus to population activity is insufficient to explain some of the most striking features of neural manifolds. First, neural manifolds can persist during sleep [26, 27, 37], and can be dissociated from sensory cues [10, 38]; meaning that the neural manifold is not a mere reflection of external stimuli. Second, the low-dimensional trajectory of the population activity observed in experiments often does not correspond to the evolution of external stimulus variables, but is instead determined by the initial condition and context [21, 31] (Fig.1E). Both findings point towards underlying recurrent circuit mechanisms [25, 37–39].

In the examples mentioned above, strong excitatory connections are found between neurons with similar preferred HD in the fly HD system [40], or between neurons with similar orientation tuning in V1 [23]. These observations are consistent with the idea that the similarity of single-neuron response properties is informative about the structure of the recurrent connectivity [12, 14, 37, 41], as the traditional notion of cortical columns postulates [42]. Conversely, the idea that the connectivity is structured by the functional properties of the neurons, such as their preferred HD in the ring model of the HD system [16], has been very influential in computational neuroscience [43–50]. However, despite these lines of experimental and theoretical works, the link between the neural manifold and the structure of the underlying recurrent connectivity has remained elusive [2, 3].

In this paper, we aim at connecting the two perspectives, by addressing the following questions. (i) Can we establish a mechanistic link between the circuit structure and the flow of the collective dynamics in the neural manifold? (ii) Is there a one-to-one mapping between the similarity space and the neural manifold? Even though their topology has been found to match in special cases (for example, a ring in the fly HD system, as illustrated in Fig.1A,D) [3, 27, 40], there is no obvious reason for why this should be true in general. (iii) Can we extract knowledge about the circuit structure from neuronal activity? While many studies have focused on extracting the neural manifold from experimental data [26, 27, 29, 30], an established method for extracting the similarity space, without explicitly referring to tuning curves or selectivity, is missing.

The results are organised as follows. First and foremost, we propose a theoretical framework providing an explicit link between the circuit structure and the collective dynamics in models of Recurrent Neural Networks (RNN). Relying on the theory of neural fields [45] and large low-rank networks [51], our framework connects functional properties of single neurons to emergent low-dimensional dynamics in a model of RNN. Maybe surprisingly, the theory predicts that there is no one-to-one relationship between the functional similarity space and the neural manifold. In particular, their respective dimensionality can be different: one can build examples where the similarity space has greater dimension than the neural manifold, and vice versa. Consequently, the same collective dynamics can result from different circuit structures. Yet, although there is no strict correspondence between the similarity space and the neural manifold, we show that they are linked through the symmetries of the connectivity. Finally, we propose an approach for extracting the circuit structure from the neuronal activity, by exploiting the similarities between the activity of single neurons. We apply it to artificial data to retrieve the circuit structure in different networks performing the same simple task. Unlike the characterisation of neuronal tuning, this

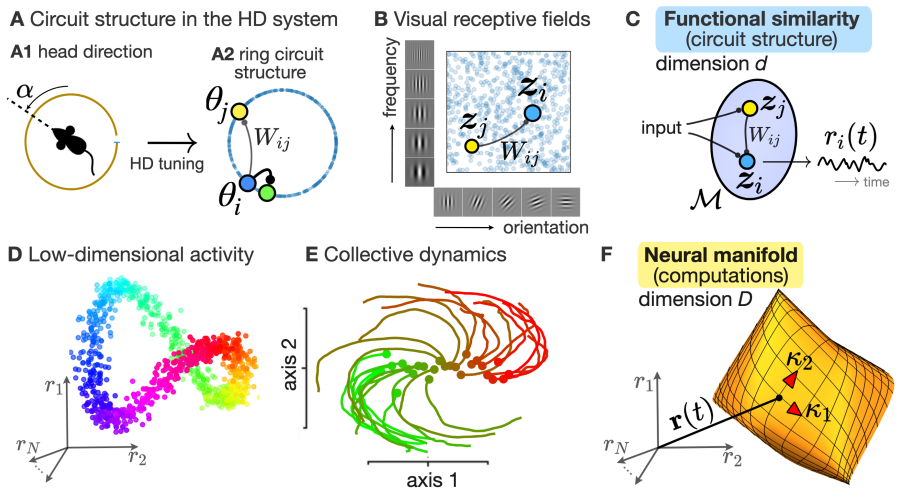

Fig. 1 *Two perspectives on neuronal activity.* **A.** The "circuit" perspective is illustrated by the ring model of the head direction (HD) system. **A1.** The animal's HD is represented by an angle α. **A2.** Each neuron in the population responds preferentially to a specific HD angle, represented as its abstract "position" (blue dots) on a one-dimensional ring [37]. This can be explained by a one-dimensional attractor model [16], where recurrent synaptic connections W_{ij} between pairs of neurons i and j are stronger between neurons that are neighbours on the ring (e.g., between the blue and green neurons). **B.** In a model of primary visual cortex, each neuron i can be characterised by the features of its receptive field (RF), e.g., orientation and frequency, which determine the neuron's "position" $\boldsymbol{z}_i$ in a 2-dimensional abstract space. In a computational model, the recurrent connectivity is determined by the respective positions of the connected neurons [52]. **C.** In this perspective, single-neuron responses are characterised by the "positions" of the neurons in an abstract space of dimension d, where neighbourhood represents functional similarity. Classical circuit models postulate that recurrent connections between pairs of neurons i and j and forward connections from external input are determined by their positions $\boldsymbol{z}_i$, $\boldsymbol{z}_j$ in this abstract space. **D.** The "neural manifold" perspective is illustrated by large-scale recordings of the neuronal activity in the rodent HD system. The time-dependent activity of a population of N cells representing the head direction angle traces out a one-dimensional ring-like manifold within the N-dimensional state-space of population activity [27]. Each point corresponds to a vector of population activity, coloured according to the HD angle α it represents. **E.** In primate motor cortex, the linear projection of the population activity onto relevant axes reveals rotational population-level trajectories during a reaching task [21]. The low-dimensional trajectory corresponding to each trial (solid lines) is mostly determined by the initial condition (dots, coloured according to their projection on axis 1). **F.** In this perspective, the joint activity of N neurons, $\boldsymbol{r}(t) \in \mathbb{R}^N$, lies on a low-dimensional *neural manifold* with intrinsic dimension $D \ll N$ (shown as a curved, $(D = 2)$-dimensional surface). The activity is parametrised by population-level latent variables $\kappa_1(t), \ldots, \kappa_D(t)$ whose dynamics reflect neural computations.
Panel (D) is inspired from [27]. Panel (E) is inspired from [21].

approach is task-agnostic, allowing us to blindly detect a circuit structure that is irrelevant to a given behavioural task.

In summary, our approach links the concepts of functional similarity of single neurons, circuit structure and neural manifold through a unifying framework encompassing field models and low-rank networks; and formalises a potential method for identifying elements of circuit structure from large-scale recordings.

Results

A unifying framework based on the theory of large RNNs.

In this section, we link the *similarity space* of single-neuron response properties with the *neural manifold* of population activity; the details of the derivations are presented in the Methods. Our framework extends recent theoretical results on large RNNs [53–55] to field models [45, 53].

Since the dynamics of large Spiking Neural Networks (SNNs) composed of Linear-Nonlinear-Poisson (LNP) neurons [56, 57] (or nonlinear Hawkes processes [58]) converges to the dynamics of equally large RNNs in the large N limit (with $1/N$ synaptic weight scaling) [55], we directly present the framework for LNP spiking neurons. In a network composed of N neurons, let the membrane potential h_i of each neuron $i = 1, \dots, N$ be a leaky, *linear* integrator (with time constant τ) of recurrent and external inputs. Each spike of a presynaptic neuron j, connected to postsynaptic neuron i, causes a stepwise change of the postsynaptic potential from h_i to $h_i + \frac{1}{N} W_{ij}$, where $\frac{1}{N} W_{ij}$ is the synaptic weight from neuron j to neuron i. The network also receives K external inputs $I_1(t), \dots, I_K(t)$, with input weights to neuron i denoted by $w_{i1}^{\text{ext}}, \dots, w_{iK}^{\text{ext}}$, respectively. The dynamics of the membrane potential h_i reads:

$$\frac{\mathrm{d}}{\mathrm{d}t} h_i(t) = -\frac{1}{\tau} h_i(t) + \frac{1}{N} \sum_{j=1}^{N} W_{ij} S_j(t) + \sum_{\nu=1}^{K} w_{i\nu}^{\text{ext}} I_\nu(t), \tag{1}$$

where $S_j(t) = \sum_s \delta(t - t_j^s)$ denotes the spike train of neuron j. The instantaneous firing rate of each neuron i is given by $\phi(h_i(t))$, where ϕ is a positive, *nonlinear* transfer function. Finally, the instantaneous firing rates $\phi(h_1(t)), \dots, \phi(h_N(t))$ define the intensity of N (conditionally independent) inhomogenous *Poisson* processes from which the spikes of each neuron are drawn.

The framework we propose relies on three assumptions. First, we assume that the connectivity matrix can be written in the form:

$$W_{ij} = \sum_{\mu=1}^{D} F_{i,\mu} G_{j,\mu} \tag{2}$$

which could result from Hebbian learning (as we discuss later) [59, 60]. This assumption is rooted in the "neural manifold" perspective, and is motivated by results on low-rank RNNs. In Eq.(2), D is a fixed number called the *rank* of the recurrent connectivity. In an RNN where the recurrent connectivity has rank D, the population activity lies in a D-dimensional neural manifold [51], in which it obeys very interpretable collective dynamics [54]. Models of low-rank networks thus enable to implement a wide range of computations – from the retrieval of patterns stored as fixed points (as in the Hopfield model) [59–61] to the approximation of arbitrary dynamical systems [35, 54].

Second, going beyond the existing theory of low-rank networks, we assume that the response properties of each neuron i can be characterized by its position $\boldsymbol{z}_i = (z_{i,1}, \dots, z_{i,d})$ in a d-dimensional *functional similarity space* $\mathcal{M}$, where neurons close to each other have similar properties. This implies that both recurrent and external connections in Eq.(1) depend on the respective positions of the neurons. This assumption is rooted in the circuit perspective described in the introduction. It is supported by evidence for recurrent mechanisms that reflect the tuning properties of the neurons [23, 37, 40], and is at the basis of many models in computational neuroscience [16, 43–50].

Lastly, in order to connect these two assumptions, we need to express how the recurrent connectivity reflects the functional similarities. Thus, we further assume that the D coefficients

$(F_{i,1}, \ldots, F_{i,D})$ and $(G_{j,1}, \ldots, G_{j,D})$ associated with neurons i and j in Eq.(2) can be expressed as functions of the neurons' respective positions $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ in the similarity space:

$$F_{i,\mu} = f_\mu(\boldsymbol{z}_i), \qquad G_{j,\mu} = g_\mu(\boldsymbol{z}_j) \tag{3}$$

where $f_1, \ldots, f_D$ and $g_1, \ldots, g_D$ are arbitrary, continuous, real-valued functions over the similarity space $\mathcal{M}$ (Fig.2). Importantly, the rank D does not need to equal the dimensionality d of the similarity space $\mathcal{M}$.

We emphasise that many known models of large low-rank networks are defined by the distribution of the coefficients $\{F_{i,\mu}, G_{i,\mu}\}$ $(\mu = 1, \ldots, D)$ associated with each neuron i across the population [51, 54, 55, 59–63]. In our framework, this distribution is parameterised by the density of neuronal positions $\boldsymbol{z}_i$ in the similarity space, and choice of the functions $\boldsymbol{f} = (f_1, \ldots, f_D)$ and $\boldsymbol{g} = (g_1, \ldots, g_D)$. In this respect, the assumption of Eq.(3) is a general formulation of the connectivity of large low-rank networks. For instance, a Hopfield model with P binary patterns [60, 61] corresponds to positions $\boldsymbol{z}_i$ uniformly distributed on the corners $\{\pm 1\}^P$ of the P-dimensional hypercube, and identity functions $f_\mu(\boldsymbol{z}_i) = g_\mu(\boldsymbol{z}_i) = z_{i,\mu}$ [64].

Taken together, these assumptions allow us to rewrite Eq.(1) as

$$\frac{\mathrm{d}}{\mathrm{d}t} h_i(t) = -\frac{1}{\tau} h_i(t) + \frac{1}{N} \sum_{j=1}^{N} w(\boldsymbol{z}_i, \boldsymbol{z}_j) S_j(t) + \sum_{\nu=1}^{K} f_\nu^{\mathrm{ext}}(\boldsymbol{z}_i) I_\nu(t), \tag{4}$$

where the functions $f_1^{\mathrm{ext}}, \ldots, f_K^{\mathrm{ext}}$ characterise the tuning of the neurons to the external inputs $I_1, \ldots, I_K$ respectively; and the connectivity kernel w is defined by Eqs.(2)-(3), i.e.

$$W_{ij} = w(\boldsymbol{z}_i, \boldsymbol{z}_j) = \sum_{\mu=1}^{D} f_\mu(\boldsymbol{z}_i) g_\mu(\boldsymbol{z}_j). \tag{5}$$

This is our first result: the general low-rank connectivity defined by Eqs.(2)-(3) corresponds to a circuit model where the connection strength between neurons i and j is given by the connectivity kernel $w(\boldsymbol{z}_i, \boldsymbol{z}_j)$. If the positions $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$ of the neurons are drawn independently, the network dynamics of Eq.(4) converge to the solution of a neural field equation in the limit of a large network [55, 65, 66], where the membrane potential of each neuron i is a function of its position $\boldsymbol{z}_i$. Therefore, we sometimes use the terms "circuit model" and "field model" interchangeably in the rest of the paper.

For the ease of exposition – and without loss of generality (as explained in the Methods) – we consider here that the rank D of the connectivity equals the dimensionality K of the input, and that the functions $(f_1, \ldots, f_D)$ and $(f_1^{\mathrm{ext}}, \ldots, f_K^{\mathrm{ext}})$ are the same. Our second, core theoretical result is that, if neuronal positions $\boldsymbol{z}_i$ are sampled independently with probability distribution $\rho(\boldsymbol{z})$, then, in the limit of a large network $(N \to \infty)$, the potentials $h_1(t), \ldots, h_N(t)$ can be written as

$$h_i(t) = \sum_{\mu=1}^{D} f_\mu(\boldsymbol{z}_i) \kappa_\mu(t) \tag{6}$$

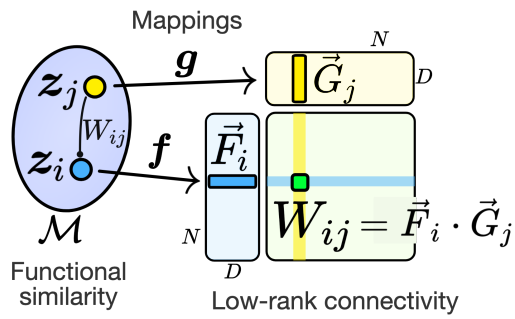

Fig. 2 *Linking the circuit structure to the low-rank connectivity.* Neurons with positions $\boldsymbol{z}_i, \boldsymbol{z}_j$ in the functional similarity space $\mathcal{M}$ (left) are mapped by two functions $\boldsymbol{f}, \boldsymbol{g} : \mathcal{M} \to \mathbb{R}^D$ to D-dimensional vectors $\vec{F}_i = \boldsymbol{f}(\boldsymbol{z}_i)$ and $\vec{G}_j = \boldsymbol{g}(\boldsymbol{z}_j)$. Each entry W_{ij} of the rank-D connectivity matrix (right) is given by the scalar product $\vec{F}_i \cdot \vec{G}_j$.

where the latent variables $\kappa_1(t), \ldots, \kappa_D(t)$ are the solution of

$$\frac{\mathrm{d}}{\mathrm{d}t}\kappa_\mu(t) = -\frac{1}{\tau}\kappa_\mu(t) + \int_{\mathcal{M}} g_\mu(\boldsymbol{z})\, \phi\Big(\sum_{\nu=1}^{D} f_\nu(\boldsymbol{z})\kappa_\nu(t)\Big)\rho(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \; + I_\mu(t). \tag{7}$$

This theoretical result tells us three things. First, Eq.(7) is deterministic. Thus, the population dynamics of large networks of stochastic LNP neurons are deterministic. Second, the population dynamics can be reduced to a nonlinear system of D ordinary differential equations (given by Eq.(7)), which means that the neural manifold is at most D-dimensional. Third, Eq.(7) expresses the dynamics of the coordinates κ_μ of the population activity in the neural manifold, and involves an integral over the similarity space $\mathcal{M}$. This provides an explicit mathematical link between the circuit structure, summarised by the embedding of the neurons in the similarity space, and the flow of the collective dynamics in the neural manifold.

As a side remark, it is worth noticing that the low-rank connectivity matrix of Eq.(5) could be the result of Hebbian learning [59, 60]. Indeed, we can interpret each function $f_\mu(\boldsymbol{z}_i)$ as the probability that neuron i with tuning properties $\boldsymbol{z}_i$ forms, in the experimental condition μ, a dendritic spine; and $g_\mu(\boldsymbol{z}_j)$ as the probability that neuron j forms a presynaptic bouton. The change of synaptic efficacy in condition μ will then be the product $\Delta W_{ij} \propto f_\mu(\boldsymbol{z}_i) g_\mu(\boldsymbol{z}_j)$: across several experimental conditions, the low-rank connectivity of Eq.(5) will build up. In a memory-related brain area, the "experimental condition" could correspond to the presentation of a novel memory item; whereas in early visual cortex the experimental condition could correspond to elongated spatial edges observed during the critical period [67].

Circuit structure and neural manifold can mismatch: examples

In the following, we highlight the insights provided by the above framework through a few examples. For the sake of simplicity, we assume in these examples that the network does not receive any external input; and that the instantaneous firing rate is a step function of the membrane potential h, with $\phi(h) = R$ for $h > 0$ and zero otherwise. Qualitatively, results do not change for other choices (Appendix B).

A one-dimensional circuit structure generates a two-dimensional neural manifold

In our first toy example, the functional properties of the neurons are summarised by their one-dimensional position $z_i \in \mathbb{R}$, which are normally distributed (Fig. 3A1). Thus, the functional similarity space $\mathcal{M} = \mathbb{R}$ is one-dimensional. If the interactions are given by $W_{ij} \propto z_i z_j$, this corresponds to a Hopfield network with a single Gaussian pattern. In this case, the collective dynamics have a single stable non-trivial fixed point, up to a symmetry degeneracy [62]. We study a variant of this model where synaptic weights are given by:

$$W_{ij} = \begin{cases} J\, z_i z_j & (z_i z_j \geq 0) \\ \cos(\alpha) J\, z_i z_j & (z_i z_j < 0) \end{cases} \tag{8}$$

with $0 < \alpha < \pi/2$; and $J > 0$ is the interaction strength. The weights of Eq.(8) can be written in the form $W_{ij} = f_1(z_i) g_1(z_j) + f_2(z_i) g_2(z_j)$ (Fig.3A2), meaning that the connectivity has rank 2. The theory thus predicts ($D = 2$)-dimensional collective dynamics, with three stable fixed points spread out in the plane and one saddle point at the origin (Fig.3B; more details in the Methods). This is confirmed by simulations of a SNN with $N = 2000$ Poisson neurons (obeying Eq.(1)) and weights given by Eq.(8). Thus, the above model generates a neural manifold of dimension $D = 2$, but has a similarity space of dimension $d = 1$.

Two different circuit structures generate identical collective dynamics

We now turn to an example of practical relevance in motor neuroscience: in a task where a monkey has to grasp a lever and cycle with its hand, the population activity lies in a neural manifold containing a self-sustained limit cycle [68]. We build here two different models that generate identical limit-cycle dynamics in a neural manifold of dimension $D = 2$.

The first circuit model is a *ring model*, in which neurons are uniformly assigned angular positions $z_i \in [0, 2\pi[$ (Fig.3C1) [16, 44]. We express the connectivity as:

$$W_{ij} = J \cos(z_i - z_j - \delta) \tag{9}$$

where $J > 0$ is the connectivity strength, and the angle $\delta \in]-\pi, \pi]$ is a phase shift parameter. For $\delta = 0$, we get a standard ring attractor model with stationary bumps of activity [44]; whereas for $\delta \neq 0$, the attractor states are travelling bumps [16]. Using the sum expansion of the cosine function [53], the connectivity can be expressed as a kernel of rank 2, where the functions in Eq.(5) are given by $f_1(z) = \cos z, f_2(z) = \sin z$ and $g_\mu(z) = J f_\mu(z + \delta)$ (for $\mu = 1, 2$).

In the second model, the neurons have two-dimensional positions $\boldsymbol{z}_i = (z_{i,1}, z_{i,2}) \in \mathbb{R}^2$ normally distributed over the two-dimensional plane (Fig.3C2). Similarly to the ring model above, the connections are maximal for an angle δ between two neurons:

$$W_{ij} = J'\, \boldsymbol{z}_i R^{(\delta)} \boldsymbol{z}_j \tag{10}$$

where $R^{(\delta)}$ is the 2×2 matrix of rotation by an angle δ in the plane; and $J' > 0$ is the connectivity strength. The connectivity of Eq.(10) has also rank 2, since it is readily expressed as a two-dimensional scalar product. Thus, the two models, with connectivity defined by Eqs.(9) and (10), generate $D = 2$-dimensional neural manifolds.

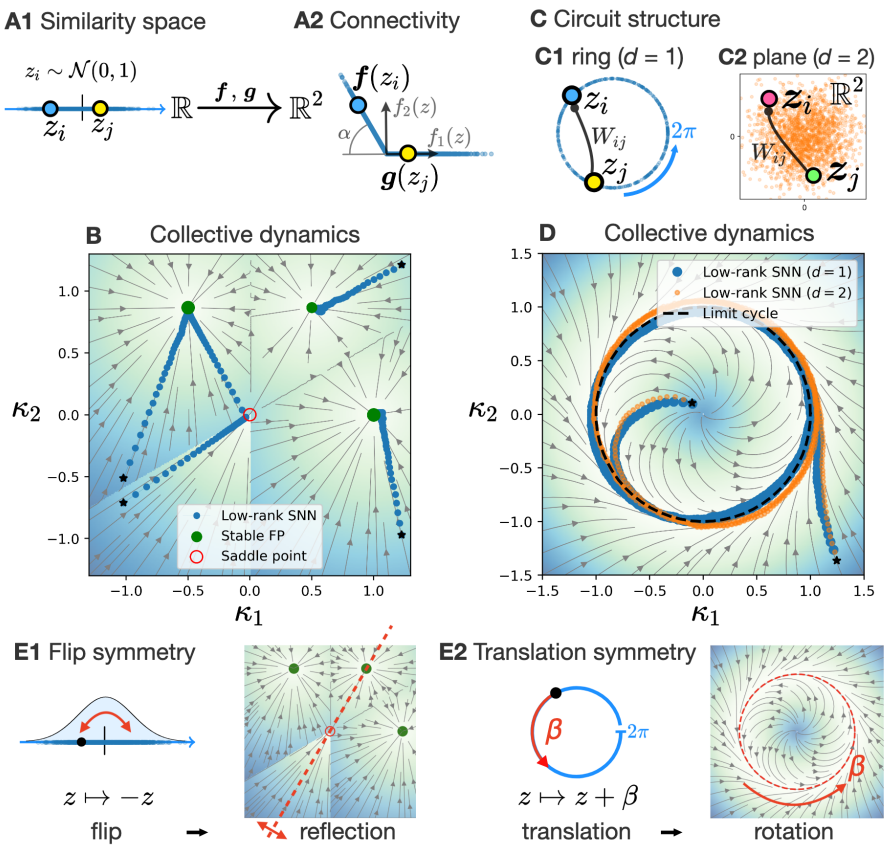

Fig. 3 *The dimension d of the functional similarity space and the dimension D of the neural manifold do not always match.* **A1.** Neurons are normally distributed over the real line: the dimension of the similarity space is $d = 1$. **A2.** The connectivity of Eq.(8) is given by 2-dimensional functions $\boldsymbol{f}$ and $\boldsymbol{g}$: the dimension of the neural manifold is $D = 2$. **B.** Flow of the collective dynamics of the circuit model in panel (A) (darker blue-shading indicates faster dynamics), given by Eq.(7). The population dynamics of an SNN with $N = 2000$ neurons (blue dots, obeying Eq.(1)) follows the flow of the collective dynamics, starting from 4 different initial conditions (black stars). The units of the axes are $k := \frac{\tau J R}{\sqrt{2\pi}}$. **C-D.** Two different circuit models can produce identical collective dynamics. **C.** Schematic of the distribution of neuron positions: **C1** uniform over the ring ($d = 1$); **C2** standard Gaussian in the plane ($d = 2$). **D.** Same as panel (B), for the circuit models in panel (C), with the connectivity of Eqs.(9)-(10). The collective dynamics generate a globally attractive limit cycle. The units of the axes are $\varrho_0 := \frac{\tau J R}{\pi} \cos\delta$. **E** The symmetries of a circuit model result in symmetries of the collective dynamics. **E1.** The symmetry w.r.t. a simultaneous flip of all neuron positions (flip symmetry) in panel (A) causes a reflection symmetry of the collective dynamics in panel (B). **E2.** Similarly, the symmetry w.r.t. simultaneous shifts of all neurons along the ring (translation symmetry) in panel (C1) causes a rotation symmetry of the collective dynamics in panel (D).
In panel (B), the SNN leaves the saddle point at the origin towards one of the attractors due to finite-size fluctuations and spike firing noise. The numerical parameters are: $\tau J R = 10$, and $\alpha = \pi/3$ in (A-B) (Eq.(8)), $\delta = \pi/10$ in (C-D) (Eqs.(9)-(10)). The low-dimensional trajectory of the SNN is obtained from Eq.(20): $\boldsymbol{\kappa}^{\mathrm{SNN}}(t) := \frac{1}{N}\sum_i M \boldsymbol{f}(z_i) h_i(t)$.

Remarkably, for $J = \sqrt{\frac{\pi}{2}}J'$ (i.e. up to a rescaling), the two models yield *identical* two-dimensional collective dynamics (more details in the Methods, Eq.(26)). The theory predicts that for any initial condition, the two latent variables that parameterise the population activity converge to a globally attractive limit cycle, which is confirmed by simulations of SNNs with $N = 2000$ Poisson neurons (Fig.3D).

We highlight two points. (i) For both the "ring" and the "planar" model, there exists a simultaneous transformation of all neuronal positions that leaves the connectivity unchanged – a translation along the ring in Eq.(9), or a rotation in the plane in Eq.(10). In turn, the flow of the collective dynamics, and therefore the set of its attractor states, exhibits a rotation symmetry, as visible in Fig.3D. The general link between the symmetries of the circuit structure and the collective dynamics is established in the next section. (ii) Neither the dimension D of the neural manifold nor the flow of the collective dynamics enable us to distinguish the ring model from the planar model. We discuss more generally how the same collective dynamics can be implemented using different circuit structures in the section "Dimensionality Mismatch". Nevertheless, we present later (section "The footprint of the circuit structure") a method that allows us to extract the circuit structure from the neuronal activity.

Symmetries of the connectivity generate symmetries of the collective dynamics

A *symmetry* of a circuit model is a simultaneous transformation of all neuronal positions that preserves the network dynamics. For instance, in the ring model defined by Eq.(9), the translation along the ring by an arbitrary (angular) distance β is a symmetry: intuitively, it means that the membrane potential dynamics of each neuron is not affected upon shifting every neuron i from angular position z_i to position $z_i + \beta$, and is a consequence of the translation invariance of the connectivity kernel of Eq.(9).

In order to define more precisely the concept of symmetry, one needs first to define the concept of *flow* of the dynamics. Consider that the membrane potential of each neuron is given, at initial time $t = 0$, by a function v_0 of its position; i.e., $h_i(0) = v_0(\boldsymbol{z}_i)$. The *flow* of the dynamics transforms the initial function v_0 into the function v_t that describes the membrane potentials at time t (i.e. $h_i(t) = v_t(\boldsymbol{z}_i)$). In other words, the flow is a mapping Φ_t such that $v_t = \Phi_t[v_0]$.

There are two ways in which a transformation S of the neuronal positions can be composed with the time evolution. Given the membrane potentials $h_i(0) = v_0(\boldsymbol{z}_i)$ at initial time $t = 0$, one can first let the system evolve up to time t, at which $h_i(t) = v_t(\boldsymbol{z}_i)$; and second, "move" each neuron i from position $\boldsymbol{z}_i$ to position $S\boldsymbol{z}_i$. The membrane potentials are then given by $h_i(t) = v_t(S\boldsymbol{z}_i) = \Phi_t[v_0](S\boldsymbol{z}_i)$. The other way is to first "move" the neurons, so that the membrane potentials at initial time are given by $h_i(0) = v_0(S\boldsymbol{z}_i)$; and second, let the system evolve up to time t. The membrane potentials are then given by $h_i(t) = (\Phi_t[v_0 \circ S])(\boldsymbol{z}_i)$ (where the circle "$\circ$" denotes the composition of functions). A symmetry of a circuit model is a transformation S of the neuronal positions that commutes with the flow of the dynamics, meaning that the two orders of composition described above are equivalent; or, in mathematical terms: $\Phi_t[v_0] \circ S = \Phi_t[v_0 \circ S]$, for any initial condition v_0 and time t. In particular, any mapping S that leaves the connectivity kernel and the distribution of the neuronal positions unchanged is a symmetry. Following the same line, a symmetry of the collective dynamics is a transformation of the latent variables $(\kappa_1, \ldots, \kappa_D)$ that commutes with the flow of the collective dynamics of Eq.(7).

We show that each symmetry S of a circuit model induces a symmetry of the associated collective dynamics in the D-dimensional neural manifold. It is a D-dimensional linear mapping $T^{(S)}$, expressed as a $D \times D$ matrix for which we give an explicit expression (Eq.(27) in the Methods). This result is further explained in the Methods and in Appendix C, in which it corresponds to Theorem 1.

For example, in the ring model above, with connectivity given by Eq.(9), each translation symmetry $S : z \mapsto z + \beta$ results in a symmetry of the collective dynamics which is given by the matrix:

$$T^{(S)} = \begin{pmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{pmatrix} \tag{11}$$

which is the matrix of rotation of an angle β in the plane (Fig.3E2). Similarly, the earlier example of a one-dimensional circuit model with connectivity given by Eq.(8) has a "flip" symmetry $S : z \mapsto -z$. The corresponding symmetry of the collective dynamics (Fig.3E1) is the reflection in two-dimensional space w.r.t. the axis spanned by the vector $(1 - \cos\alpha, \sin\alpha)$.

This result provides a link between the similarity space and the neural manifold: the symmetries of a circuit model, expressed as transformations of the neuronal positions in the similarity space, generate symmetries of the collective dynamics in the neural manifold.

Dimensionality mismatch

Despite this link provided by the symmetries, the above example of two models generating the same limit-cycle dynamics (Fig.3C,D) illustrates that the functional similarity space can have different dimension than the neural manifold, and that several circuit models can generate the same collective dynamics. Here, we discuss two possible mechanisms for this mismatch.

Heterogeneity as a mechanism of dimensionality mismatch.

First, the dimension d of the functional similarity space can be increased by additional parameters of single-neuron properties that parameterise heterogeneity across the population. As a starting point, consider our "planar" limit-cycle model above with a two-dimensional circuit structure (Eq.(10)). In this example, neuronal positions are normally distributed in the two-dimensional plane. Thus, the position $\boldsymbol{z}_i \in \mathbb{R}^2$ of any neuron i can be written in polar coordinates, with an angle φ_i distributed uniformly over the ring $[0, 2\pi[$, and a radius ξ_i (distributed following the χ_2-distribution). We interpret the radius ξ_i as a single-neuron property that parameterises heterogeneity of both the transfer functions and connectivity across neurons, compared to a "pure" ring model (in which $\xi_i = 1$ for all neurons). Yet, this heterogeneity does not affect the collective dynamics of the ring model (up to a rescaling; see Appendix B.2). Along the same line of argument, additional single-neuron properties that parameterise heterogeneity across the population can be incorporated into any circuit model, thus extending the dimensionality d of the similarity space, without affecting the collective dynamics. Simple examples include the multiplication of the outgoing weights of each neuron by a random factor, so as to describe variations of synapse strength; and heterogeneous maximum firing rates across neurons (see Appendix D).

Several universal approximators.

Dimensionality mismatch can also arise from the fact that several circuit models with different circuit structures can equivalently implement any target collective dynamics. We distinguish two types of circuit models that are "universal approximators". On the one hand, a Gaussian mixture low-rank network of rank D can approximate arbitrary collective dynamics in a D-dimensional

neural manifold [54]. This model has a functional similarity space of dimension $d = 2D$ (Appendix E.1). On the other hand, field models with a one-dimensional similarity space can in principle also approximate arbitrary collective dynamics, if the functions f_μ, g_μ in Eq.(5) can be chosen arbitrarily with no requirement of "smoothness" – analogous to the universal approximation property of feed-forward neural networks (Appendix E.2).

The footprint of the circuit structure

We conclude from the previous sections that there is no unique relationship between the collective dynamics of the population activity in the neural manifold, and the circuit structure, reflected by the embedding of the neurons in the functional similarity space. Yet, the collective dynamics form a specific, low-dimensional summary of the activity of hundreds of neurons. Even if the circuit structure provably can not be extracted from the collective dynamics, can it be recovered by analysing the neuronal activity? In this section, we propose an approach to this problem, that we illustrate in a model of contextual decision-making.

Individual statistics provide embeddings of the neurons' positions.

A recording of the activity of N neurons at $T \gg d$ time points can be represented as a $N \times T$ matrix $\mathbf{r}$, whose entries $\mathbf{r}_{i,k} = r_i(t_k)$ correspond to the estimated firing rate of neuron i at time t_k (Fig.4A). Each *column* $\mathbf{r}_{:,k} \in \mathbb{R}^N$ of this matrix represents the population activity at time t_k as a N-dimensional vector. If sufficiently many time points are recorded, dimensionality reduction can be robustly performed on the resulting N-dimensional point-cloud, enabling the decoding of the collective dynamics from the population activity [26–31].

On the other hand, in a circuit (or field) model, the firing rate $r_i(t) = r(t, \boldsymbol{z}_i)$ of any neuron i depends on its position $\boldsymbol{z}_i$ in the functional similarity space. Therefore, each *row* $\mathbf{r}_{i,:}$ of the activity matrix $\mathbf{r}$ represents the activity of a single neuron i as a T-dimensional vector that depends on its position $\boldsymbol{z}_i$:

$$\mathbf{r}_{i,:} = (r(t_1, \boldsymbol{z}_i), r(t_2, \boldsymbol{z}_i), \ldots, r(t_T, \boldsymbol{z}_i)) \in \mathbb{R}^T \tag{12}$$

Since neuronal positions $\boldsymbol{z}_i$ reflect the similarity of single-neuron responses, the T-dimensional vectors associated with neurons at neighbouring positions $\boldsymbol{z}_i, \boldsymbol{z}_j$ will in turn be similar. If sufficiently many neurons are recorded, the resulting set of T-dimensional vectors will form an embedding of the d-dimensional similarity space.

As an illustration, we consider the two previous example models that generate limit-cycle collective dynamics in a two-dimensional neural manifold (Eqs.(9) and (10)), where we assume a sigmoidal voltage-to-rate function $\phi(h) = R\,(\tanh(h) + 1)/2$. Although in this case, the collective dynamics of the two models are not strictly identical, they remain similar (Appendix B.2), so that the activity of the corresponding SNNs follow similar low-dimensional trajectories in the neural manifold (Fig.4A,B). Yet, the value of the firing rate of any neuron i at three different time points t_1, t_2, t_3 yields a 3-dimensional embedding of the neuron, given by $(r_i(t_1), r_i(t_2), r_i(t_3))$, which summarises its activity. The point-cloud formed by the embeddings of all the neurons in the population reveals the functional similarity space of each model, allowing us to distinguish between the two different circuit structures (Fig.4C).

Instead of directly using the estimated firing rate of each neuron at T different time points, this reasoning can be extended to other sets of T neuron-specific statistics that summarise its activity: for instance, the z-scored spike counts of each given neuron i across T different time bins, or the contribution ("loadings") of T principal components of the population activity to the activity of

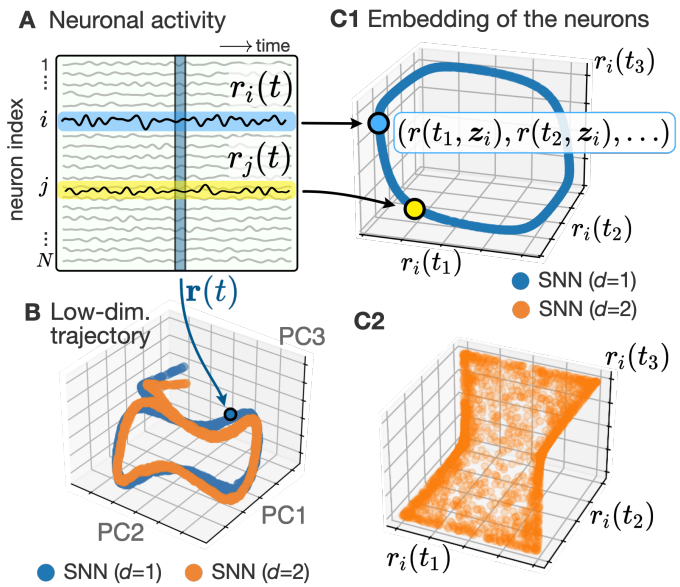

Fig. 4 *The functional similarity space can be extracted from the neuronal activity.* **A.** In a circuit model, each *column* of the population activity, i.e., the *momentary* firing rate vector $\boldsymbol{r}(t) \in \mathbb{R}^N$, lies on the neural manifold; whereas each *row* corresponds to a given neuron's time-dependent firing rate $r_i(t) = r(t, \boldsymbol{z}_i)$, parameterised by its position $\boldsymbol{z}_i$ in the functional similarity space. **B.** The population activity of two qualitatively different RNNs of $N = 2000$ neurons follow similar limit-cycle dynamics in the neural manifold, even though their respective functional similarity spaces have different dimension d (see Fig 3C,D). We show here the trajectory of the three first principal components (PCs) of the population activity, to which a rotation was applied to superpose the trajectories of the two models. **C1.** Estimates of the firing rate $(r_i(t_1), r_i(t_2), r_i(t_3))$ of each neuron i at three different time points yields a three-dimensional embedding of the neuron. For the first model, the point-cloud formed by the embeddings of all the neurons in the population forms a ring, revealing the functional similarity space characterising the underlying ring circuit structure (Fig. 3C1). **C2.** In constrast, for the second model, the set of all three-dimensional embeddings lies on a curved two-dimensional surface, consistent with the two-dimensional similarity space of Fig. 3C2.

each neuron i (as illustrated in the next section). Such statistics, recorded across many neurons, will reveal the functional similarity space, embedded in T dimensions – provided that these statistics preserve functional similarities, by being similar for two neurons that are close to each other in the similarity space, and dissimilar if neurons are further apart.

Two circuit models of context-dependent decision-making

We illustrate the extraction of the circuit structure with a simple model of context-dependent decision-making [31]. During each trial, the network receives a sensory input consisting of two independent signals (named "color" and "motion") that each fluctuates in time around some fixed positive or negative mean value; as well as a binary context input which indicates the relevant signal: a value of $+1$ (respectively -1) indicates the "color" (resp. "motion") context (Fig.5A). The task consists in indicating, in the "color" context, whether the "color" signal is positive or negative at trial onset – and similarly for the "motion" context. This task can be solved by two-dimensional collective dynamics [51].

We reformulate the task in our framework. In a circuit model, the input weights, i.e., the contribution of the sensory and contextual signals to the external input received by each neuron i,

are parameterised by functions of the neuron's position $\boldsymbol{z}_i$, as expressed by Eq.(4). Similarly, the output is a one-dimensional readout of the spiking activity, with weights determined by the neurons' positions. We study two circuit models with different circuit structures that both solve the task. For both models, in between two trials, the population activity decays towards a baseline state; but in the "color" context, symmetric attractive fixed points appear along a "color"-relevant axis, and similarly in the "motion" context (Fig.5B). We call two different SNNs of $N = 3000$ neurons, corresponding to the two circuit models, the "blue" and the "orange" networks. We aim to identify the circuit structure underlying each of the two networks from the neuronal activity.

Extracting the circuit structure from the neuronal activity.

First, two-dimensional trajectories corresponding to the population activity of each SNN were obtained by projecting the N-dimensional spiking activity onto "color" and "motion"-relevant axes [31]. Upon receiving the same sensory and context inputs, the low-dimensional trajectories and the choice readouts of the two SNNs are similar (Fig.5C,D,E).

On the other hand, statistical indicators of the activity of individual neurons, based on spike counts within three different time windows, yield a 3-dimensional embedding of each neuron. For both models, the resulting point-cloud lies in a two-dimensional subspace (Fig.5F). Using persistent homology to characterise the topology of these point-clouds [26, 27, 69, 70], the ring topology of the functional similarity space is confirmed by a persistent single ring-feature for the "blue" model, but not the "orange" one (Fig. 5G). The identification of the functional similarity space from statistical features is correct and reflects the underlying circuit structure of each model (Fig.5H). The "blue" network is a ring model, with neuronal positions uniformly distributed over the ring; whereas the "orange" network has a 2-dimensional circuit structure, with neurons uniformly distributed in a square.

Extracting a task-irrelevant circuit structure

In the previous example, the difference between the circuit structure underlying each of the two "blue" and "orange" SNNs could have been detected by inspecting the tuning of each neuron to the sensory input features. Suppose that we assign to each neuron a two-dimensional "tuning vector", that characterises its response to the two sensory input features "color" and "motion". In both models, individual neurons are tuned to mixtures of these sensory features ("mixed selectivity" [14]). For the "orange" network, the tuning of each neuron to the two features is uncorrelated. Yet in the "blue" network, the tuning of each neuron to these two features is correlated: the corresponding two-dimensional tuning vector lies on a circle in $\mathbb{R}^2$.

However, our approach is task-agnostic, allowing us to detect a circuit structure that is irrelevant for the task at hand. As an illustration, we consider a third model. Starting from the same ring circuit structure as the "blue" SNN, the position of each neuron i, characterised by its position z_i on the ring, is "enriched" by a task-irrelevant feature $\xi_i = \pm 1$ (chosen independently for each neuron i, with equal probability of being $+1$ or -1). The connection strength between pairs of neurons i and j is in turn modified by the addition of a task-irrelevant component, given by $K\,\xi_i\xi_j$, with arbitrary strength $K \geq 0$ (Fig.6A). The original "blue" network corresponds to a coupling strength $K = 0$.

Since this additional feature is irrelevant the task at hand, the additional circuit structure cannot be detected through the characterisation of neuronal tuning. Moreover, the two-dimensional projection of the population activity onto "color"- and "motion"-relevant axes yields similar trajectories for the original model (for which $K = 0$) and the "enriched" model with $K > 0$ (Fig.6B).

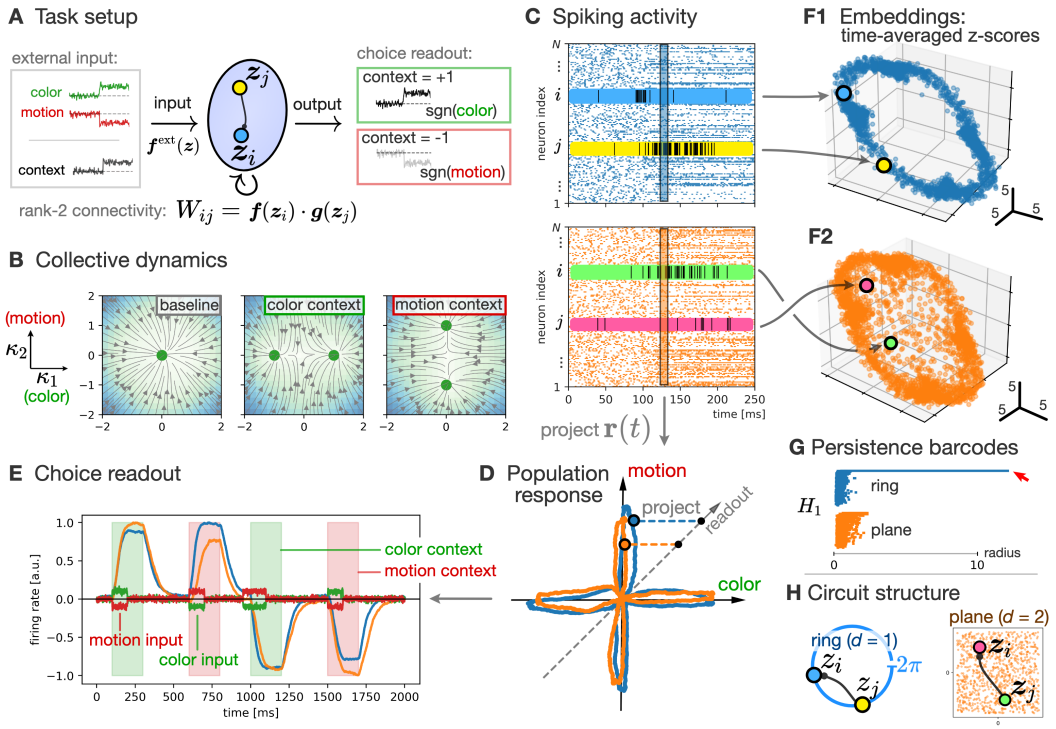

Fig. 5 *Context-dependent decision-making in two SNNs with different circuit structures.* **A.** During each trial, the network receives an external input consisting of two sensory features (named "color" and "motion"), each fluctuating around a fixed mean value that can be positive or negative, and a binary context feature. The task of the network is to indicate the sign of the sensory feature that is specified by the context ("color" for the context being $+1$, and "motion" for -1). Input, recurrent, and output synaptic weights are determined by functions of the neurons' positions $\boldsymbol{z}$ in the functional similarity space. **B.** The task is solved by two-dimensional collective dynamics, where the context input generates attractors (green dots) along the axis that corresponds to the relevant sensory input feature. **C.** Spike trains of 150 neurons from two SNNs ("blue" SNN, top; "orange" SNN, bottom; $N = 3000$ neurons each). Spike trains of example neurons i and j are highlighted. **D.** Low-dimensional trajectory of the two SNNs during 4 trials, obtained by projecting the N-dimensional spiking activity $\boldsymbol{r}(t)$ onto "color"- and "motion"-relevant axes. The choice readout is the projection of the population activity on a diagonal readout axis which is a linear combination of the "color" and "motion" axes. **E.** Time-course of the external inputs (green and red traces for the "color" and "motion" features, respectively) and normalised choice readout of the two networks (blue and orange traces) during the 4 trials. The context input takes value $+1$ (respectively -1) in the "color" (resp. "motion") context, indicated by green (resp. red) shadings, and 0 in between trials. Both networks correctly solve the task. **F.** 3-dimensional embeddings of the neurons, generated by the time-average of each neuron's z-scored activity within three different time intervals. Both point-clouds in panels (F1) and (F2) lie within a 2-dimensional linear subspace (99% variance explained by 2 first principal components). **G.** Persistence barcodes of the ring features H_1 of the simplicial complexes associated with the point-clouds of panels (F1)-(F2). The ring topology of the point-cloud in (F1) is characterised by the persistent ring feature (long bar indicated with red arrow). **H.** The functional similarity space extracted from the neural activity (a ring with $d = 1$ for the "blue" SNN, and a plane with $d = 2$ for the "orange" SNN) corresponds to the underlying circuit structure.

Yet, the set of three-dimensional embeddings of all the neurons in the population, based on a Principal Component Analysis (PCA) of the population activity, yields a point-cloud that reveals the underlying circuit structure. For $K = 0$, one recovers the ring structure of the original "blue" model

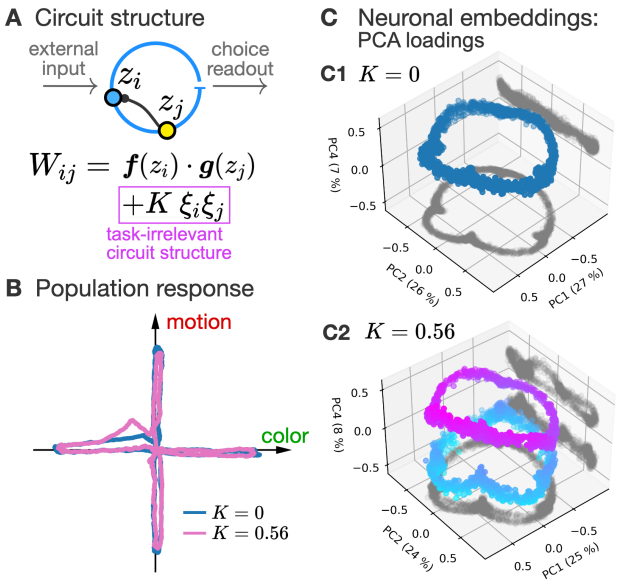

Fig. 6 *Extracting a task-irrelevant circuit structure in a model of context-dependent decision-making.* **A.** The task is the same as depicted in Fig.5A. It is solved by two SNNs of $N = 3000$ neurons each. The first one corresponds to a ring model (the "blue" SNN of Fig.5). In the second one, each neuron's position z_i on the ring is enriched with an additional task-irrelevant binary feature $\xi_i = \pm 1$, distributed independently across neurons. The connectivity is modified by an additional task-irrelevant component of strength $K \geq 0$. **B.** Low-dimensional trajectory of two SNNs (one with $K = 0$, corresponding to the original "blue" model; and one with an additional task-irrelevant connectivity component of strength $K = 0.56 > 0$) during 4 trials during which they received the same external input, obtained by projecting the N-dimensional spiking activity onto task-relevant axes (defined as in Fig.5C-D). **C.** 3-dimensional embeddings of the neurons, corresponding to the contribution of the first, second and fourth components of the PCA decomposition of the population activity to the spiking activity of each neuron i. The percentage of variance of the population activity explained by each component of the PCA is indicated on the corresponding axis. **C1.** The ring circuit structure of the original SNN is recovered for $K = 0$. **C2.** For a strength $K = 0.56$, the point-cloud forms two disjoint rings, revealing the task-irrelevant circuit structure of the second SNN. Points are coloured according to their projection on the vertical axis. Points on the upper (respectively, lower) ring correspond to neurons with a task-irrelevant feature $\xi_i = +1$ (resp., $\xi_i = -1$).

(Fig.6C1); whereas for $K > 0$, the point-cloud forms two disjoint rings, revealing the topology of the functional similarity space.

Discussion

To address the relationship between the functional similarity space, assumed to reflect the structure of cortical circuits, and the neural manifold, assumed to characterise computations via the collective dynamics of the population activity, we formulated three specific questions. As an answer to the first question, we proposed a theoretical framework that provides an explicit link between the connectivity of a circuit model and the flow of the collective dynamics in the neural manifold. Our framework relies on the key idea that in the limit of large networks, both low-rank RNNs and classical circuit models, where the connectivity is structured by the functional properties of the neurons can be interpreted in the general framework of neural field models. In passing, we thus

addressed the open problem of the relationship between classical circuit models and low-rank RNNs [3].

The second question needs to be answered negatively: there is no unique relationship between the functional similarity space and the neural manifold. As a consequence, different circuit structures can yield equivalent collective dynamics. However, we showed that the symmetries of the connectivity are inherited by the collective dynamics. Moreover, we highlighted heterogeneity across the population as a potential cause for the mismatch between the dimensionality of the functional similarity space and of the neural manifold.

As an answer to the third question, we proposed a method for extracting the topology of the functional similarity space from the population activity, independently of the collective dynamics. In contrast with the analysis of neuronal tuning or selectivity, this approach is task-agnostic, which allowed us to extract a circuit structure that was irrelevant to the task at hand. In conjunction with task- and behaviour-informed methods [29], such "unsupervised tuning discovery" could enable the extraction of intricate features modulating complex neuronal responses in higher cortical areas.

While our model opens the door to a unified perspective on the link between circuit structure, neuronal activity, and behaviour, it does not incorporate important features of more realistic models of biological neural networks. First, in our model, recurrent connectivity and neuronal properties are dominated by a homogeneous structural principle whereas cortical connectivity appears to be, at least partially, random [71]; moreover, neuronal responses exhibit a degree of disorder [14, 31, 72]. These aspects have been incorporated into several earlier models and can lead to chaotic or chaos-like activity [51, 59, 73]. The generalisation of our theory to the case of mixed structured and random connectivity, following [51], is left for future work.

Second, the neurons in our model are linear-nonlinear-Poisson neurons, which do not account for refractoriness. Our theory can be partly extended to models of renewal neurons (such as the Spike Response Model or Integrate-and-Fire models [74, 75]). In this case, the collective dynamics can no longer be described as an ODE. Yet, in the regime of asynchronous firing, each neuron's firing rate can be approximated by a non-linear function of its synaptic input [76], so that our formalism can be applied to the description of the fixed points of the asynchronous population activity.

Third, cortical networks exhibit a balance between excitation and inhibition [77, 78]. This has been accounted for in models of balanced networks [79–81], including spatially structured networks [82]. The extension of our formalism to spatially structured balanced networks remains an open problem.

Finally, we did not apply our method for the extraction of the circuit structure to experimental data. The question of whether such structure could be observed in experimental recordings is, at this stage, still open. A positive answer would suggest that, even in a population where neurons exhibit mixed selectivity and seemingly heterogeneous responses, the degree of heterogeneity in the population is reducible to a low-dimensional description, and that the neuronal activity can be described by field models. Our work contributes to link this classic approach in neuroscience, based on the classification of neurons by receptive fields and response similarities, to a modern population-level approach that leverages the statistical power of large-scale neural recordings.

References

[1] Urai, A.E., Doiron, B., Leifer, A.M., Churchland, A.K.: Large-scale neural recordings call for new insights to link brain and behavior. Nature Neuroscience **25**(1), 11–19 (2022)

[2] Barack, D.L., Krakauer, J.W.: Two views on the cognitive brain. Nature Reviews Neuroscience **22**(6), 359–371 (2021)

[3] Langdon, C., Genkin, M., Engel, T.A.: A unifying perspective on neural manifolds and circuits for cognition. Nature Reviews Neuroscience **24**(6), 363–377 (2023)

[4] Hubel, D.H., Wiesel, T.N.: Receptive fields of single neurones in the cat's striate cortex. The Journal of Physiology **148**(3), 574–591 (1959)

[5] J.P. Rauschecker, M.H. B. Tian: Processing of complex sounds in the macaque nonprimary auditory cortex. Science **268**, 111–114 (1995)

[6] Georgopoulos, A.P., Schwartz, A., Kettner, R.E.: Neuronal population coding of movement direction. Science **233**, 1416–1419 (1986)

[7] McNaughton, B.L., Barnes, C.A., O'Keefe, J.: The contributions of position, direction, and velocity to single unit activity in the hippocampus of freely-moving rats. Experimental Brain Research **52**, 41–49 (1983)

[8] Moser, E.I., Moser, M.-B., McNaughton, B.L.: Spatial representation in the hippocampal formation: a history. Nature Neuroscience **20**(11), 1448–1464 (2017)

[9] Taube, J., Muller, R., JB Ranck, J.: Head-direction cells recorded from the postsubiculum in freely moving rats. 1. description and quantitative analysis. Journal of Neuroscience **10**(2), 420–435 (1990)

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

[11] Doeller, C.F., Barry, C., Burgess, N.: Evidence for grid cells in a human memory network. Nature **463**(7281), 657–661 (2010)

[12] Hirokawa, J., Vaughan, A., Masset, P., Ott, T., Kepecs, A.: Frontal cortex neuron types categorically encode single decision variables. Nature **576**(7787), 446–451 (2019)

[13] Raposo, D., Kaufman, M.T., Churchland, A.K.: A category-free neural population supports evolving demands during decision-making. Nature Neuroscience **17**(12), 1784–1792 (2014)

[14] Rigotti, M., Barak, O., Warden, M.R., Wang, X.-J., Daw, N.D., Miller, E.K., Fusi, S.: The importance of mixed selectivity in complex cognitive tasks. Nature **497**(7451), 585–590 (2013)

[15] Taube, J.S.: The head direction signal: Origins and sensory-motor integration. Annual Review of Neuroscience **30**(Volume 30, 2007), 181–207 (2007)

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

[17] Bednar, J.A., Wilson, S.P.: Cortical maps. The Neuroscientist **22**, 604–617 (2016)

[18] Musall, S., Kaufman, M.T., Juavinett, A.L., Gluf, S., Churchland, A.K.: Single-trial neural dynamics are dominated by richly varied movements. Nature Neuroscience **22**(10), 1677–1686 (2019)

[19] Bugeon, S., Duffield, J., Dipoppa, M., Ritoux, A., Prankerd, I., Nicoloutsopoulos, D., Orme, D., Shinn, M., Peng, H., Forrest, H., Viduolyte, A., Reddy, C.B., Isogai, Y., Carandini, M., Harris, K.D.: A transcriptomic axis predicts state modulation of cortical interneurons. Nature **607**(7918), 330–338 (2022)

[20] Aoi, M.C., Mante, V., Pillow, J.W.: Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. Nature Neuroscience **23**(11), 1410–1420 (2020)

[21] Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., Shenoy, K.V.: Neural population dynamics during reaching. Nature **487**(7405), 51–56 (2012)

[22] Bonhoeffer, T., Grinvald, A.: lso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. Nature **353**, 429–431 (1991)

[23] Bosking, W.H., Zhang, Y., Schofield, B., Fitzpatrick, D.: Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. J Neurosci **17**(6), 2112–2127 (1997)

[24] Saxena, S., Cunningham, J.P.: Towards the neural population doctrine. Current Opinion in Neurobiology **55**, 103–111 (2019)

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

[26] Gardner, R.J., Hermansen, E., Pachitariu, M., Burak, Y., Baas, N.A., Dunn, B.A., Moser, M.-B., Moser, E.I.: Toroidal topology of population activity in grid cells. Nature **602**(7895), 123–128 (2022)

[27] Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A., Fiete, I.: The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. Nature Neuroscience **22**(9), 1512–1520 (2019)

[28] Gallego, J.A., Perich, M.G., Miller, L.E., Solla, S.A.: Neural manifolds for the control of movement. Neuron **94**(5), 978–984 (2017)

[29] Schneider, S., Lee, J.H., Mathis, M.W.: Learnable latent embeddings for joint behavioural and neural analysis. Nature **617**(7960), 360–368 (2023)

[30] Jensen, K.T., Kao, T.-C., Tripodi, M., Hennequin, G.: Manifold gplvms for discovering non-euclidean latent structure in neural data. Advances in Neural Information Processing Systems (2020)

[31] Mante, V., Sussillo, D., Shenoy, K.V., Newsome, W.T.: Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature **503**(7474), 78–84 (2013)

[32] Vyas, S., Golub, M.D., Sussillo, D., Shenoy, K.V.: Computation through neural population dynamics. Annual Review of Neuroscience **43**(1), 249–275 (2020)

[33] Shenoy, K.V., Sahani, M., Churchland, M.M.: Cortical control of arm movements: A dynamical systems perspective. Annual Review of Neuroscience **36**(1), 337–359 (2013)

[34] Safaie, M., Chang, J.C., Park, J., Miller, L.E., Dudman, J.T., Perich, M.G., Gallego, J.A.: Preserved neural dynamics across animals performing similar behaviour. Nature **623**(7988), 765–771 (2023)

[35] DePasquale, B., Sussillo, D., Abbott, L.F., Churchland, M.M.: The centrality of population-level factors to network computation is demonstrated by a versatile approach for training spiking networks. Neuron **111**(5), 631–64910 (2023)

[36] Kriegeskorte, N., Wei, X.-X.: Neural tuning and representational geometry. Nature Reviews Neuroscience **22**(11), 703–718 (2021)

[37] Peyrache, A., Lacroix, M.M., Petersen, P.C., Buzsáki, G.: Internally organized mechanisms of the head direction sense. Nature Neuroscience **18**(4), 569–575 (2015)

[38] Waaga, T., Agmon, H., Normand, V.A., Nagelhus, A., Gardner, R.J., Moser, M.-B., Moser, E.I., Burak, Y.: Grid-cell modules remain coordinated when neural activity is dissociated from external sensory cues. Neuron **110**(11), 1843–18566 (2022)

[39] Burak, Y.: Spatial coding and attractor dynamics of grid cells in the entorhinal cortex. Current Opinion in Neurobiology **25**, 169–175 (2014)

[40] Kim, S.S., Rouault, H., Druckmann, S., Jayaraman, V.: Ring attractor dynamics in the drosophila central brain. Science **356**(6340), 849–853 (2017)

[41] Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F., Ostojic, S.: The role of population structure in computations through neural dynamics. Nature Neuroscience **25**(6), 783–794 (2022)

[42] Mountcastle, V.B.: The columnar organization of the neocortex. Brain **120**(4), 701–722 (1997)

[43] Amari, S.-i.: Dynamics of pattern formation in lateral-inhibition type neural fields. Biological Cybernetics **27**(2), 77–87 (1977)

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

[45] Wilson, H.R., Cowan, J.D.: A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. Kybernetik **13**(2), 55–80 (1973)

[46] Burak, Y., Fiete, I.R.: Accurate Path Integration in Continuous Attractor Network Models of Grid Cells. PLOS Computational Biology **5**(2), 1000291 (2009)

[47] McNaughton, B.L., Battaglia, F.P., Jensen, O., Moser, E.I., Moser, M.-B.: Path integration and the neural basis of the 'cognitive map'. Nature Reviews Neuroscience **7**(8), 663–678 (2006)

[48] Coombes, S., Beim Graben, P., Potthast, R., Wright, J.: Neural Fields: Theory and Applications. Springer, Heidelberg (2014)

[49] Licurgo de Almeida, Marco Idiart, John E. Lisman: The Input–Output Transformation of the Hippocampal Granule Cells: From Grid Cells to Place Fields. The Journal of Neuroscience **29**(23), 7504 (2009)

[50] Zhou, D., Rangan, A.V., McLaughlin, D.W., Cai, D.: Spatiotemporal dynamics of neuronal population response in the primary visual cortex. Proceedings of the National Academy of Sciences **110**(23), 9517–9522 (2013)

[51] Mastrogiuseppe, F., Ostojic, S.: Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. Neuron **99**(3), 609–62329 (2018)

[52] Cai, D., Rangan, A., McLoughlin, D.W.: Architectural and synaptic mechanisms underlying coherent spontaneous activity in V1. Proceedings of the National Academy of Sciences **102**, 5868–5873 (2005)

[53] Veltz, R., Faugeras, O.: Local/global analysis of the stationary solutions of some neural field equations. SIAM Journal on Applied Dynamical Systems **9**(3), 954–998 (2010)

[54] Beiran, M., Dubreuil, A., Valente, A., Mastrogiuseppe, F., Ostojic, S.: Shaping Dynamics With Multiple Populations in Low-Rank Recurrent Networks. Neural Computation **33**(6), 1572–1615 (2021)

[55] Schmutz, V., Brea, J., Gerstner, W.: Emergent rate-based dynamics in duplicate-free populations of spiking neurons. arXiv:2303.05174 (2024)

[56] Chichilnisky, E.J.: A simple white noise analysis of neuronal light responses. Network **12**(2), 199–213 (2001)

[57] Gerstner, W., Kistler, W.M., Naud, R., Paninski, L.: Neuronal Dynamics. From Single Neurons to Networks and Models of Cognition. Cambridge University Press, Cambridge (2014)

[58] Brémaud, P., Massoulié, L.: Stability of nonlinear Hawkes processes. The Annals of Probability **24**(3), 1563–1588 (1996)

[59] Pereira, U., Brunel, N.: Attractor dynamics in networks with learning rules inferred from in vivo data. Neuron **99**(1), 227–2384 (2018)

[60] Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences **79**(8), 2554–2558 (1982)

[61] Hopfield, J.J.: 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)

[62] Amit, D.J., Gutfreund, H., Sompolinsky, H.: Spin-glass models of neural networks. Phys. Rev. A **32**, 1007–1018 (1985)

[63] Schuessler, F., Dubreuil, A., Mastrogiuseppe, F., Ostojic, S., Barak, O.: Dynamics of random recurrent networks with correlated low-rank structure. Phys. Rev. Res. **2**, 013111 (2020)

[64] Herz, A., Sulzer, B., Kühn, R., Hemmen, J.L.: Hebbian learning reconsidered: Representation of static and dynamic objects in associative neural nets. Biological Cybernetics **60**(6), 457–467 (1989)

[65] Gerstner, W.: Time structure of the activity in neural network models. Phys. Rev. E **51**, 738–758 (1995)

[66] Chevallier, J., Duarte, A., Löcherbach, E., Ost, G.: Mean field limits for nonlinear spatially extended hawkes processes with exponential memory kernels. Stochastic Processes and their Applications **129**(1), 1–27 (2019)

[67] Espinosa, J.S., Stryker, M.P.: Development and plasticity of the primary visual cortex. Neuron **75**(2), 230–249 (2012)

[68] Saxena, S., Russo, A.A., Cunningham, J., Churchland, M.M.: Motor cortex activity across movement speeds is predicted by network-level strategies for generating muscle activity. eLife **11**, 67620 (2022)

[69] Rybakken, E., Baas, N., Dunn, B.: Decoding of Neural Data Using Cohomological Feature Extraction. Neural Computation **31**(1), 68–93 (2019)

[70] Ghrist, R.: Barcodes: the persistent topology of data. Bull. Amer. Math. Soc. (N.S.) **45**(1), 61–75 (2008)

[71] Harris, K.D., Mrsic-Flogel, T.D.: Cortical connectivity and sensory coding. Nature **503**(7474), 51–58 (2013)

[72] Churchland, M.M., Shenoy, K.V.: Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. Journal of Neurophysiology **97**(6), 4235–4257 (2007)

[73] Sompolinsky, H., Crisanti, A., Sommers, H.J.: Chaos in random neural networks. Phys. Rev. Lett. **61**, 259–262 (1988)

[74] Gerstner, W.: Population dynamics of spiking neurons: Fast transients, asynchronous states, and locking **12**(1), 43–89 (2000)

[75] Spiridon, M., Gerstner, W.: Effect of lateral connections on the accuracy of the population code for a network of spiking neurons. Network: Computation in Neural Systems **12**(4), 409 (2001)

[76] Ostojic, S., Brunel, N.: From spiking neuron models to linear-nonlinear models. PLOS Computational Biology **7**(1), 1–16 (2011)

[77] Shu, Y., Hasenstaub, A., McCormick, D.A.: Turning on and off recurrent balanced cortical activity. Nature **423**(6937), 288–293 (2003)

[78] Haider, B., Duque, A., Hasenstaub, A.R., McCormick, D.A.: Neocortical network activity in vivo is generated through a dynamic balance of excitation and inhibition. Journal of Neuroscience **26**(17), 4535–4545 (2006)

[79] Brunel, N.: Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. Journal of Computational Neuroscience **8**(3), 183–208 (2000)

[80] Renart, A., Rocha, J., Bartho, P., Hollender, L., Parga, N., Reyes, A., Harris, K.D.: The asynchronous state in cortical circuits. Science **327**(5965), 587–590 (2010)

[81] Litwin-Kumar, A., Doiron, B.: Slow dynamics and high variability in balanced cortical networks with clustered connections. Nature Neuroscience **15**(11), 1498–1505 (2012)

[82] Rosenbaum, R., Doiron, B.: Balanced networks of spiking neurons with spatially dependent recurrent connections. Phys. Rev. X **4**, 021039 (2014)

[83] Fournier, N., Löcherbach, E.: On a toy model of interacting neurons. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques **52**(4), 1844–1876 (2016)

[84] Faugeras, O., Grimbert, F., Slotine, J.-J.: Absolute stability and complete synchronization in a class of neural fields models. SIAM Journal on Applied Mathematics **69**(1), 205–250 (2008)

[85] Renardy, M., Rogers, R.C.: Operator Theory, 2nd edn., pp. 228–282. Springer, New York (NY) (2004)

[86] Conway, J.B.: A Course in Functional Analysis, pp. 42–43267. Springer, New York (NY) (2007)

[87] Cybenko, G.: Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems **2**(4), 303–314 (1989)

Methods

A unifying framework: field model of low-rank networks

Network model.

We turn here to a more precise mathematical formulation. We consider a model of Recurrent Neural Network (RNN) composed of N linear-nonlinear-Poisson neurons [56]. The spike train $S_i(t)$ of each neuron $i = 1, \ldots, N$ is generated by an inhomogeneous Poisson process with intensity $\phi(h_i(t))$, where $h_i(t)$ represents the neuron's membrane potential, and $\phi : \mathbb{R} \to \mathbb{R}^+$ is a non-linear bounded

voltage-to-rate function. The potential $h_i(t)$ is a leaky integrator of the synaptic input; which, for the ease of exposition, we first consider as coming only from recurrent connections. The dynamics of the potential thus follow:

$$\frac{\mathrm{d}}{\mathrm{d}t}h_i(t) = -\frac{1}{\tau}h_i(t) + \frac{1}{N}\sum_{\substack{j=1\\ j\neq i}}^{N} W_{ij}S_j(t) \tag{13}$$

with given initial condition $h_i(0)$. Here, $\tau > 0$ is the membrane time constant; and the synaptic weights, scaled as $1/N$, are given by a matrix of rank $D \in \mathbb{N}$ fixed:

$$W_{ij} = \sum_{\mu=1}^{D} F_{i,\mu}G_{j,\mu} = \vec{F}_i \cdot \vec{G}_j \tag{14}$$

where the dot denotes the scalar product in $\mathbb{R}^D$. Each entry W_{ij} of the low-rank connectivity matrix corresponds to the scalar product of two D-dimensional "connectivity vectors", $\vec{F}_i = (F_{i,1}, \ldots, F_{i,D})$ and $\vec{G}_i = (G_{i,1}, \ldots, G_{i,D})$, associated with neurons i and j (see Fig.2). As discussed in the Results section, Eq.(14) is a general formulation of the constraint of a low-rank connectivity; specific models of low-rank networks are essentially built by specifying the distribution of these vectors across the population [51, 54, 55, 59–63]. In what follows, we introduce a general model for the distribution of the connectivity vectors. The resulting class of RNNs encompasses most low-rank models encountered in the literature. Moreover, for fixed rank D, as $N \to \infty$, it corresponds to a field model over a space whose dimension d is not necessarily related to the rank D of the connectivity, i.e. to the dimension of the neural manifold.

A general model of low-rank connectivity.

As we are interested in the limit of large N, let's start from the simple observation that, if one wants to add a neuron labelled i^* to the network, one has to specify its connectivity, through the associated vectors $\vec{F}_{i^*}$ and $\vec{G}_{i^*}$, as well as its initial condition $h_{i^*}(0)$. It is commonly assumed that these features are independent across neurons: intuitively, it means one can add a neuron i^* to the network without knowing the $\{\vec{F}_i, \vec{G}_i, h_i(0)\}_{i=1}^N$ associated with all other neurons. However, for a given neuron i, the components of $\vec{F}_i$ and $\vec{G}_i$ might be correlated. For the new neuron i^*, we thus model these features as being determined by some individual properties $\boldsymbol{z}_{i^*}$ of the neuron, sampled from a given distribution.

Formally, we represent $\boldsymbol{z}_{i^*}$ by the neuron's *position* on a smooth manifold $\mathcal{M}$ of intrinsic dimension d, drawn at random according to a probability distribution $\rho(\mathrm{d}\boldsymbol{z})$. Given the position $\boldsymbol{z}_i$ of any neuron i, the associated connectivity vectors $\vec{F}_i$, $\vec{G}_i$ are determined by mappings $\boldsymbol{f}, \boldsymbol{g} : \mathcal{M} \to \mathbb{R}^D$:

$$F_{i,\mu} = f_\mu(\boldsymbol{z}_i),\ G_{i,\mu} = g_\mu(\boldsymbol{z}_i) \quad (\mu = 1, \ldots, D) \tag{15}$$

with f_μ, g_μ the μ-th components of $\boldsymbol{f}$, $\boldsymbol{g}$. This is illustrated on Fig.2. Analogously, the initial condition is determined by a map $h_0 : \mathcal{M} \to \mathbb{R}$, according to $h_i(0) = h_0(\boldsymbol{z}_i)$. Eqs.(14),(15) define a kernel $w(\boldsymbol{z}, \boldsymbol{z}') = \boldsymbol{f}(\boldsymbol{z}) \cdot \boldsymbol{g}(\boldsymbol{z}')$ such that the recurrent connectivity is given by $W_{ij} = w(\boldsymbol{z}_i, \boldsymbol{z}_j)$. The manifold $\mathcal{M}$ thus represents the functional similarity space, in which the neurons' positions

determine the connectivity. The distribution of the neurons over this space, described by $\rho(\mathrm{d}\boldsymbol{z})$, defines a scalar product between functions over $\mathcal{M}$:

$$\langle \varphi, \psi \rangle_\rho := \int_{\mathcal{M}} \rho(\mathrm{d}\boldsymbol{z}) \varphi(\boldsymbol{z}) \psi(\boldsymbol{z}) \tag{16}$$

In the following, we assume that the components f_μ, g_μ ($\mu = 1, \ldots, D$) of the connectivity mappings, as well as the initial condition h_0, are continuous and belong to the vector space of ρ-square-integrable functions over $\mathcal{M}$. This last condition defines a Hilbert space $\mathcal{H}$ equipped with the ρ-scalar product $\langle \cdot, \cdot \rangle_\rho$.

Corresponding field model.

As $N \to \infty$, the RNN defined by Eqs. (13,14,15) is described by a field model: the membrane potential at time t of a neuron at position $\boldsymbol{z}$ is given by the neural field $v_t(\boldsymbol{z})$, which obeys the *field equation*

$$\partial_t v_t(\boldsymbol{z}) = -\frac{1}{\tau} v_t(\boldsymbol{z}) + \int_{\mathcal{M}} \rho(\mathrm{d}\boldsymbol{z}')\, \boldsymbol{f}(\boldsymbol{z}) \cdot \boldsymbol{g}(\boldsymbol{z}')\, \phi(v_t(\boldsymbol{z}')) \tag{17}$$

with initial condition $v_0(\boldsymbol{z}) = h_0(\boldsymbol{z})$. It is the continuum version of Eq.(13) [57, 65], where the Poisson noise from the spikes vanishes [55, 66]; i.e., $S_i(t)$ was replaced by the expected instantaneous rate $\phi(h_i(t))$. Under suitable regularity and boundedness assumptions concerning the transfer function ϕ, the convergence as $N \to \infty$ of each membrane potential $h_i(t)$ to $v_t(\boldsymbol{z}_i)$ is guaranteed, with rate $1/\sqrt{N}$. It follows that neurons become pairwise independent [66, 83].

Low-dimensional dynamics.

Analogously to a rank-D RNN [54], the dynamics of the neural field can be written as a D-dimensional system [53], as presented here in a simple fashion. The field v_t, at any time t, is an element of the Hilbert space $\mathcal{H}$ (provided that $h_0, \{f_\mu\}_\mu \subset \mathcal{H}$). The field equation (17) can thus be written as an ODE in $\mathcal{H}$: [84]

$$\partial_t v_t = -\frac{1}{\tau} v_t + \sum_{\mu=1}^{D} f_\mu \langle g_\mu, \phi \circ v_t \rangle_\rho \tag{18}$$

where $(\circ)$ denotes the composition of functions, and $\langle \cdot, \cdot \rangle_\rho$ the ρ-scalar product of Eq.(16). The field v_t undergoes an exponential decay, except in the linear subspace $\mathcal{F} := \mathrm{span}\{f_\mu\} \subset \mathcal{H}$. It can thus be written as a linear combination of the basis functions $\{f_\mu\}$, whose coefficients $\kappa_\mu(t)$ are the "latent variables":

$$v_t(\boldsymbol{z}) = \sum_{\mu=1}^{D} f_\mu(\boldsymbol{z}) \kappa_\mu(t) \; + e^{-t/\tau} h_0^\perp(\boldsymbol{z}) \tag{19}$$

where $h_0^\perp = h_0 - \boldsymbol{f} \cdot \boldsymbol{\kappa}(0)$ is the projection of the initial condition h_0 onto the orthogonal complement of $\mathcal{F}$ in $\mathcal{H}$. By projection of Eq.(19) onto each basis function f_μ, the latent variables can be written in vector form as:

$$\boldsymbol{\kappa}(t) = M \langle \boldsymbol{f}, v_t \rangle_\rho \tag{20}$$

where we introduced the inverse correlation matrix M, defined by $(M^{-1})_{\mu\nu} = \langle f_\mu, f_\nu \rangle_\rho$ (assuming without loss of generality that the functions $\{f_\mu\}$ are linearly independent in $\mathcal{H}$). Eqs. (18-19-20) yield D-dimensional collective dynamics for $\boldsymbol{\kappa}$. It can be written compactly:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\kappa}(t) = -\frac{1}{\tau}\boldsymbol{\kappa}(t) + \langle \boldsymbol{g}, \phi \circ \big(\boldsymbol{f} \cdot \boldsymbol{\kappa}(t)\big)\rangle_\rho \tag{21}$$

under the assumption that $h_0^\perp = 0$. This equation for the collective dynamics (and its more general expression, Eq.(43) in the Appendix) is the central ingredient of our theory. The derivation is given in Appendix A.

Neural manifold and attractor states.

Eq.(19) implies that the firing rate of each neuron i, given by $r_i(t) = \phi(h_i(t)) = \phi(v_t(\boldsymbol{z}_i))$, is parametrised by the latent variables $\boldsymbol{\kappa}$; and so is the N-dimensional network activity $\boldsymbol{r}(t)$. Besides, $\boldsymbol{r}(t)$ exponentially decays towards the D-dimensional neural manifold

$$\{\phi(F \cdot \boldsymbol{k}) \mid \boldsymbol{k} \in \mathbb{R}^D\} \subset \mathbb{R}^N \tag{22}$$

generated by the set of all possible latent variables; where $F \in \mathbb{R}^{N\times D}$ denotes the matrix of connectivity vectors $\vec{F}_i$ given by Eq.(15), and ϕ is applied element-wise. In the absence of external input, the latent variables $\boldsymbol{\kappa}(t)$ will converge to the set of attractor states of the collective dynamics of Eq.(21), which specifically depends on the considered model. In the Results, we give concrete examples, in which attractor states are either discrete fixed points or limit cycles.

Generalisations: external input and arbitrary connectivity kernel.

This derivation can be readily extended to the case of a network receiving a low-dimensional external input. We assume that the external input is composed of K features $I_\nu(t)$ ($\nu = 1, \dots, K$), and that the tuning of any neuron i to each input feature I_ν is determined by a corresponding function $f_\nu^{\mathrm{ext}}(\boldsymbol{z}_i)$ of its position $\boldsymbol{z}_i$. The external input it receives can be written in the form:

$$I_i^{\mathrm{ext}}(t) = \sum_{\nu=1}^{K} f_\nu^{\mathrm{ext}}(\boldsymbol{z}_i) I_\nu(t) \tag{23}$$

This simply results in a drive to the collective dynamics of Eq.(21), while expanding their dimensionality – at most up to $D + K$. The corresponding general expression of the collective dynamics is given by Eq.(43) in Appendix A.2.

The derivation of the collective dynamics can also be generalised to the case of field models with arbitrary square-integrable interaction kernel over $\mathcal{M}^2$. In this case, the collective dynamics are not guaranteed to be finite-dimensional; but the field can always be well approximated by a low-rank model by a truncation of the dimension [53]. Both generalisations are given in Appendix A.

Latent dimension vs. field dimension.

There is no assumption in our framework regarding any link between the dimension D of the neural manifold (which is determined by the rank of the connectivity) on one side, and the dimension d

of the functional similarity space on the other side. Moreover, it is not clear whether the circuit structure restricts the range of collective dynamics that can be implemented in an RNN, expressed by Eq.(21). These are the points we address in the Results.

Circuit structure and neural manifold can mismatch: examples

A one-dimensional circuit structure generates two-dimensional collective dynamics

In our first example, neuronal positions are normally distributed over the real line; i.e. $\mathcal{M} = \mathbb{R}$ and $z_i \sim \rho = \mathcal{N}(0,1)$ i.i.d. for $i = 1, \dots, N$ (Fig.3A1). The interactions are given by:

$$W_{ij} = \begin{cases} J\, z_i z_j & (z_i z_j \geq 0) \\ \cos(\alpha) J\, z_i z_j & (z_i z_j < 0) \end{cases} \tag{24}$$

with $0 < \alpha < \pi/2$. This can be written as the 2-dimensional scalar product $W_{ij} = \boldsymbol{f}(z_i) \cdot \boldsymbol{g}(z_j)$, with:

$$\boldsymbol{f}(z) = \begin{pmatrix} [z]_+ + \cos(\alpha)[z]_- \\ -\sin(\alpha)[z]_- \end{pmatrix} \in \mathbb{R}^2 \quad ; \quad \boldsymbol{g} = J\boldsymbol{f} \tag{25}$$

where $[z]_\pm = z \cdot \mathbb{1}_{\{\mathrm{sgn}(z) = \pm 1\}}$ denotes the positive or negative part of z. This is illustrated on Fig.3A2. The corresponding collective dynamics resulting from Eq.(21) and displayed in Fig.3B are derived in Appendix B.1.

Two different circuit structures generate identical limit-cycle dynamics

Let $J = \sqrt{\frac{\pi}{2}} J'$. The two models with connectivity given by Eqs.(9)-(10) obey *identical* collective dynamics: for the polar coordinates $(\varrho(t), \theta(t))$ of the latent variable $\boldsymbol{\kappa}(t) \in \mathbb{R}^2$, Eq.(21) yields

$$\dot{\varrho} = -\frac{1}{\tau}\varrho + \frac{J}{\pi} R \cos\delta \quad ; \quad \varrho\dot{\theta} = \frac{J}{\pi} R \sin\delta \tag{26}$$

for *both* models (derivation in Appendix B.2). For any initial condition, the latent variables converge to a globally attractive limit cycle with radius $\varrho_0 = \frac{J}{\pi}\tau R \cos\delta$ and constant angular velocity $\dot{\theta} = \frac{1}{\tau}\tan\delta$.

Symmetries

In mathematical terms, a *symmetry* of the field model is a mapping $S : \mathcal{M} \to \mathcal{M}$ over the space of neuronal positions for which the field dynamics of Eq.(17) are equivariant: i.e., such that given any solution v_t to the dynamics, $v_t \circ S$ is also a solution. Similarly, a symmetry of the collective dynamics is a mapping T over $\mathbb{R}^D$ which is an equivariance of Eq.(21), i.e., for any solution $\boldsymbol{\kappa}(t)$ to Eq.(21), $T\boldsymbol{\kappa}(t)$ is also a solution. We show the following (see Theorem 1 in Appendix C):

1. Let S be (i) a symmetry of the field model, and (ii) under which the D-dimensional space $\mathcal{F} = \mathrm{span}\{f_\mu\}_{\mu=1}^D$ containing the recurrent input is invariant. This is expressed by the condition: $f_\nu \circ S \in \mathcal{F}$, for all $\nu = 1, \dots, D$. Then, a symmetry of the collective dynamics is given by the

linear mapping $T^{(S)}$, defined by the $D \times D$ matrix

$$T^{(S)}_{\mu\nu} = \sum_{\sigma} M_{\mu\sigma} \langle f_\sigma, f_\nu \circ S \rangle_\rho \tag{27}$$

with M the inverse correlation matrix, defined by $(M^{-1})_{\mu\nu} = \langle f_\mu, f_\nu \rangle_\rho$.

2. If S is a bijection for which both the connectivity kernel w and the distribution of the positions ρ are invariant, i.e. $w(Sz, Sz') = w(z, z')$ and $\rho \circ S(dz) = \rho(dz)$, then both requirements (i) and (ii) of point 1. above are satisfied.

Furthermore, the mapping $\mathcal{T} : S \mapsto T^{(S)}$ defines the action on $\mathbb{R}^D$ of the symmetry group G of the field model, which will also be a symmetry group of the collective dynamics. Yet, the order of composition of group elements is reversed: for any elements S, S' of G acting on $\mathcal{M}$, the action on $\mathbb{R}^D$ of their composition $S \circ S'$ is given by: $T^{(S \circ S')} = T^{(S')} T^{(S)}$. Therefore, a "right" group action on $\mathcal{M}$ becomes a "left" group action on $\mathbb{R}^D$: see Corollary 1 in Appendix C. This will have an effect if the symmetry group G is non-commutative; like, for example, the group of rotations in $\mathbb{R}^3$.

Model of context-dependent decision-making

For both models, the contribution of the sensory and contextual input features (respectively: I_1, "color", I_2, "motion", and I_c, "context") to the external input received by each neuron i is parameterised by functions $f_1^{\mathrm{ext}}(z_i), f_2^{\mathrm{ext}}(z_i)$ and $f_\mathrm{c}(z_i)$, respectively. We impose the recurrent and the sensory input to be colinear by setting $f_\mu^{\mathrm{ext}} = f_\mu$ ($\mu = 1, 2$). The total external input received by each neuron i reads:

$$I(t, z_i) = \boldsymbol{f}(z_i) \cdot \boldsymbol{I}(t) + J_\mathrm{c} f_\mathrm{c}(z_i) I_\mathrm{c}(t) \tag{28}$$

where $\boldsymbol{I} = (I_1, I_2)$ is the sensory input, and J_c is an adjustable input strength parameter. During each trial, $I_1(t)$ and $I_2(t)$ fluctuate around a fixed mean $\pm\bar{I}$, which is either positive or negative with fixed magnitude $\bar{I}$: i.e. $I_\mu(t) = \pm\bar{I} + \sigma^2 \xi_\mu(t)$, where $\xi_\mu(t)$ is a Gaussian white noise and σ^2 is the variance of the fluctuations. The context input takes binary values $I_\mathrm{c}(t) = \pm 1$.

Since the maps f_1, f_2, f_c determine the contribution of the sensory features I_1, I_2 and context I_c to the external input each neuron receives, they can be understood as encoding the neuron's selectivity to these features.

Collective dynamics

We assume symmetric couplings, i.e. $\boldsymbol{g} = J\boldsymbol{f}$; and arbitrary voltage-to-rate function ϕ. For a stationary context input $I_\mathrm{c}(t) = \pm 1$, the collective dynamics write (from Eq.(43)):

$$\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{\kappa}(t) = -\frac{1}{\tau} \boldsymbol{\kappa}(t) + J \langle \boldsymbol{f}, \phi \circ (\boldsymbol{f} \cdot \boldsymbol{\kappa}(t) + \tau J_\mathrm{c} f_\mathrm{c} I_\mathrm{c}) \rangle_\rho + \boldsymbol{I}(t) \tag{29}$$

The feature maps $\boldsymbol{f}(z)$, the context selectivity $f_\mathrm{c}(z)$ and the strength parameters J, J_c can be conveniently chosen such that in the absence of external input ($I_1 = I_2 = I_\mathrm{c} = 0$), these dynamics admit no non-trivial fixed point; but in the "color" context ($I_\mathrm{c} = +1$), symmetric attractive fixed points appear along the "color" axis at $\boldsymbol{\kappa} = (\pm k, 0)$, with $k > 0$. Similarly, the "motion" context ($I_\mathrm{c} = -1$) yields attractive fixed points along the "motion" axis at $\boldsymbol{\kappa} = (0, \pm k)$, as illustrated in Fig.5B.

This is achieved by setting $f_c = |f_1| - |f_2|$. In the "color" context $I_c = +1$, neurons that are more selective to the first input feature ("color"), i.e. for which $f_1(\boldsymbol{z}_i) > f_2(\boldsymbol{z}_i)$, are excited; while those that are more selective to the second input feature ("motion") are inhibited; and reciprocally in the "motion" context $I_c = -1$.

At the beginning of each trial, the context input is switched on to $I_c = \pm 1$, indicating the context. The population activity is then driven towards one of the fixed points by the relevant component of the sensory input (I_1, "color" if $I_c = +1$, or I_2, "motion" if $I_c = -1$), while ignoring the irrelevant component; and it relaxes towards the baseline state when the context input is switched off, being ready for a new trial.

Choice readout

The goal of the task is to match the output of the network with the sign of the relevant input feature, which is specified by the context. The output is a one-dimensional linear readout of the activity, with output weights $f_{\text{out}}(\boldsymbol{z}_i)$:

$$\text{readout}(t) = \frac{1}{N} \sum_{i=1}^{N} f_{\text{out}}(\boldsymbol{z}_i) S_i(t) \tag{30}$$

with S_i the spike train of neuron i. The map of output weights is set to $f_{\text{out}} = f_1 + f_2$. Thus, the readout is positive for the two fixed points that correspond to a positive sensory input ($(+k, 0)$ in the "color" context if $I_1 > 0$, or $(0, +k)$ in the "motion" context if $I_2 > 0$), and similarly, negative for negative sensory inputs.

Low-dimensional trajectory of the SNN

Low-dimensional trajectories of the SNNs in Fig.5D were obtained by projecting the low-passed spiking activity onto "color" and "motion"-related axes, defined by $\{f_\mu(\boldsymbol{z}_i)\}_{i=1}^N$ (with $\mu = 1$ for the "color" and $\mu = 2$ for the "motion" axis). Since $f_{\text{out}} = f_1 + f_2$, the "choice" readout of Eq.(30) is equivalent to the projection of this 2-dimensional population response on a diagonal readout axis (Fig.5D,E).

Appendix

Note: in the Appendix, we abandon the boldface notation $\boldsymbol{z}$ to denote neuronal positions. Instead, the notation z still represents a d-dimensional variable in the functional similarity space $\mathcal{M}$.

A Collective dynamics

A.1 Derivation of the collective dynamics

We start from eq.(18), rewritten here for convenience:

$$\partial_t v_t = -\frac{1}{\tau} v_t + \sum_{\mu=1}^{D} f_\mu \langle g_\mu, \phi \circ v_t \rangle_\rho \tag{31}$$

The recurrent input belongs to the linear subspace $\mathcal{F} = \text{span}\{f_\mu\} \subset \mathcal{H}$. We can decompose the field v_t as its projection onto $\mathcal{F}$, plus a component $v_t^\perp$ belonging to its orthogonal complement $\mathcal{F}^\perp$:

$$v_t = \sum_\mu f_\mu \kappa_\mu(t) + v_t^\perp \tag{32}$$

which implies:

$$\partial_t v_t = \sum_\mu f_\mu \dot{\kappa}_\mu(t) + \partial_t v_t^\perp \tag{33}$$

where the dot ($\dot{}$) denotes the time derivative. Together, equations (31)-(32) yield:

$$\partial_t v_t = \sum_\mu f_\mu \left(-\frac{1}{\tau} \kappa_\mu(t) + \langle g_\mu, \phi \circ v_t \rangle_\rho \right) - \frac{1}{\tau} v_t^\perp \tag{34}$$

As both $v_t^\perp$ and $\partial_t v_t^\perp$ are orthogonal to all of the f_μ, we get by identification of eqs.(33) and (34):

$$\dot{\kappa}_\mu(t) = -\frac{1}{\tau} \kappa_\mu(t) + \langle g_\mu, \phi \circ v_t \rangle_\rho \tag{35}$$

$$\partial_t v_t^\perp = -\frac{1}{\tau} v_t^\perp \tag{36}$$

Thus, $v_t^\perp = e^{-t/\tau} h_0^\perp$, so that eq.(32) yields eq.(19). It allows replacing v_t in eq.(35) to obtain, in vector form:

$$\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{\kappa}(t) = -\frac{1}{\tau} \boldsymbol{\kappa}(t) + \langle \boldsymbol{g}, \phi \circ \left(\boldsymbol{f} \cdot \boldsymbol{\kappa}(t) + e^{-t/\tau} h_0^\perp \right) \rangle_\rho \tag{37}$$

which yields eq.(21) if we assume $h_0 \in \mathcal{F}$.

Eventually, the expression of the latent variables $\boldsymbol{\kappa}$ can be obtained by projecting eq.(32) on $\mathcal{F}$: $\langle f_\nu, v_t \rangle_\rho = \sum_\mu \langle f_\nu, f_\mu \rangle_\rho \, \kappa_\mu$; or in vector form: $\langle \boldsymbol{f}, v_t \rangle_\rho = C \boldsymbol{\kappa}$, with the correlation matrix $C_{\nu\mu} := \langle f_\nu, f_\mu \rangle_\rho$. We can assume without loss of generality that the set of functions $\{f_\mu\}$ are linearly independent in $\mathcal{H}$. If that is not the case, a new basis of $\mathcal{F}$ with fewer elements can always be chosen (see also the next section A.2). Thus, the matrix C has full rank D, and is thus invertible, yielding eq.(20):

$$\boldsymbol{\kappa}(t) = C^{-1} \langle \boldsymbol{f}, v_t \rangle_\rho \tag{38}$$

A.2 Including external inputs

Following the arguments of section 2, assume that the network receives an external input composed of K features $I_\nu(t)$ ($\nu = 1, \dots, K$), to which the tuning of any neuron i is determined by a corresponding function $f_\nu^{\text{ext}}(z_i)$ of its position z_i. The external input received by any neuron can be written: $I^{\text{ext}}(t, z_i) = \sum_{\nu=1}^{K} f_\nu^{\text{ext}}(z_i) I_\nu(t)$. The total (recurrent + external) input thus now belongs to the linear subspace $\tilde{\mathcal{F}} \subset \mathcal{H}$ spanned by $\{f_\mu\}_{\mu=1}^{D} \cup \{f_\nu^{\text{ext}}\}_{\nu=1}^{K}$, of dimension $\tilde{D} \leq D + K$. Let $\{\tilde{f}_\alpha\}_{\alpha=1}^{\tilde{D}}$ be a basis of $\tilde{\mathcal{F}}$. Since any $f_\mu \in \mathcal{F}$ is a linear combination of the $\{\tilde{f}_\alpha\}$, then for each α, there exists a linear combination $\tilde{g}_\alpha$ of the feature maps $\{g_\mu\}$ such that

$$w(z, z') = \sum_{\mu=1}^{D} f_\mu(z) g_\mu(z') = \sum_{\alpha=1}^{\tilde{D}} \tilde{f}_\alpha(z) \tilde{g}_\alpha(z') \tag{39}$$

and similarly, there exists a linear combination $\tilde{I}_\alpha$ of the input features $\{I_\nu\}$ such that

$$I^{\text{ext}}(t, z) = \sum_{\nu=1}^{K} f_\nu^{\text{ext}}(z) I_\nu(t) = \sum_{\alpha=1}^{\tilde{D}} \tilde{f}_\alpha(z) \tilde{I}_\alpha(t) \tag{40}$$

Without loss of generality, we can thus write the total synaptic input received at position z as:

$$I(t, z) = \sum_{\alpha=1}^{\tilde{D}} \tilde{f}_\alpha(z) \Big(\int_{\mathcal{M}} \tilde{g}_\alpha(z') \phi(v_t(z')) \rho(\mathrm{d}z) \, + \tilde{I}_\alpha(t) \Big) \tag{41}$$

From now on, we omit the tildes ($\tilde{\circ}$) for readability. The field equation in the case of a low-dimensional external input can be written:

$$\partial_t v_t = -\frac{1}{\tau} v_t + \sum_\alpha f_\alpha \Big(\langle g_\alpha, \phi \circ v_t \rangle_\rho + I_\alpha(t) \Big) \tag{42}$$

Repeating the derivation of sec.A.1 gives the general expression of the collective dynamics:

$$\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{\kappa}(t) = -\frac{1}{\tau} \boldsymbol{\kappa}(t) + \langle \boldsymbol{g}, \phi \circ \Big(\boldsymbol{f} \cdot \boldsymbol{\kappa}(t) + e^{-t/\tau} h_0^{\perp} \Big) \rangle_\rho \, + \boldsymbol{I}(t) \tag{43}$$

where $\boldsymbol{I}(t) = (I_1(t), \dots, I_D(t))$ is the vector of external input features.

A.3 Generalisation to arbitrary square-integrable kernels

We turn now to the case of field models where the connectivity is determined by a kernel $w(z, z')$ over $\mathcal{M}^2$ which is only assumed to be in $L^2(\rho \times \rho)$, i.e.

$$\|w\|_{2,\rho\times\rho}^2 := \int_{\mathcal{M}^2} \rho(\mathrm{d}z) \rho(\mathrm{d}z') w(z, z')^2 < \infty \tag{44}$$

yielding the general field equation:

$$\begin{aligned}\partial_t v_t(z) &= -\frac{1}{\tau} v_t(z) + \int_{\mathcal{M}} \rho(dz') w(z,z') \phi\left(v_t(z')\right) \\ &= -\frac{1}{\tau} v_t + K_w[\phi \circ v_t]\end{aligned} \tag{45}$$

where $K_w : \mathcal{H} \to \mathcal{H}$ is the Hilbert-Schmidt integral operator associated with the kernel w: [85]

$$K_w[\varphi](z) := \int_{\mathcal{M}} \rho(\mathrm{d}z') w(z,z') \varphi(z') \tag{46}$$

Since K_w is a Hilbert-Schmidt operator, it can always be decomposed as:

$$K_w[\varphi] = \sum_{\alpha \in \mathbb{N}} \lambda_\alpha x_\alpha \langle y_\alpha, \varphi \rangle_\rho \tag{47}$$

where the sets of left and right eigenvectors $\{x_\alpha\}, \{y_\alpha\}$ are orthonormal families in $\mathcal{H}$. The l^2-norm of the series of singular values $\lambda_\alpha \geq 0$ is finite, and is equal to the Hilbert-Schmidt norm of K_w, i.e. $\sum_{\alpha \in \mathbb{N}} \lambda_\alpha^2 = \|w\|_{2,\rho\times\rho}^2 < \infty$ [85, 86]. The dynamics of general fields, eq.(45), have thus a very similar structure to that of low-rank fields, eq.(17) – except (precisely) that the rank of K_w is not necessarily finite. In particular, one can still write the associated collective dynamics for the projections $\kappa_\alpha(t) = \langle x_\alpha, u_t \rangle_\rho$, following the line of sec.A.1:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} \kappa_\alpha(t) &= -\frac{1}{\tau} \kappa_\alpha(t) + \lambda_\alpha \langle y_\alpha, \phi \circ v_t \rangle_\rho \qquad (\alpha \in I) \\ v_t &= \sum_{\alpha \in I} x_\alpha \kappa_\alpha(t) + e^{-t/\tau} v_0^\perp\end{aligned} \tag{48}$$

where $I = \{\alpha \mid \lambda_\alpha > 0\} \subseteq \mathbb{N}$, and $v_0^\perp = v_0 - \sum_\alpha x_\alpha \kappa_\alpha(0)$. From this point of view, a low-rank field is only characterised by a finite set of non-zero singular values $\lambda_\alpha > 0$. The dynamics of any arbitrary field can be approximated using a truncation of the sum in eq.(47) [53], which corresponds to a model of low-rank network according to our framework (see eq.(15)).

B Examples of collective dynamics

B.1 Fixed points in a generalised Hopfield model

We consider here the connectivity mappings given by eq.(25), which read:

$$\boldsymbol{f}(z) = \begin{pmatrix} [z]_+ + \cos\alpha [z]_- \\ -\sin\alpha [z]_- \end{pmatrix} = \begin{pmatrix} 1 & \cos\alpha \\ 0 & -\sin\alpha \end{pmatrix} \begin{pmatrix} [z]_+ \\ [z]_- \end{pmatrix} =: \mathbf{A}\, \tilde{\boldsymbol{f}}(z) \ ; \ \boldsymbol{g} = J\boldsymbol{f} \tag{49}$$

with $\tilde{\boldsymbol{f}}(z) = ([z]_+, [z]_-)$, $\mathbf{A}$ the 2×2 matrix above, and $[z]_\pm = z \cdot \mathbb{1}_{\{\mathrm{sgn}(z)=\pm 1\}}$. Let the neurons be normally distributed over the real line, i.e. $z_i \sim \mathcal{N}(0,1)$. The collective dynamics of eq.(21) write:

$$(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau})\boldsymbol{\kappa} = \int_{\mathbb{R}} Dz\, J\mathbf{A}\tilde{\boldsymbol{f}}(z)\phi(\boldsymbol{\kappa}\mathbf{A}\tilde{\boldsymbol{f}}(z)) \tag{50}$$

with Dz the standard Gaussian measure; and we assume for now a general transfer function ϕ. We make the following change of variable: $\boldsymbol{\lambda} = \mathbf{A}^T\boldsymbol{\kappa}$. Noticing that $(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau})\boldsymbol{\lambda} = \mathbf{A}^T(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau})\boldsymbol{\kappa}$, we have:

$$\begin{aligned}(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau})\boldsymbol{\lambda} &= \int_{\mathbb{R}} Dz\, J\mathbf{A}^T\mathbf{A}\tilde{\boldsymbol{f}}(z)\phi(\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{f}}(z)) \\ &= J\mathbf{A}^T\mathbf{A}\int_{\mathbb{R}} Dz \begin{pmatrix}[z]_+\\ [z]_-\end{pmatrix}\phi(\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{f}}(z))\end{aligned} \tag{51}$$

Let r, θ be the polar coordinates of $\boldsymbol{\lambda}$. From the expression of $(\lambda_1, \lambda_2) = r(\cos\theta, \sin\theta)$, we get:

$$\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{f}}(z) = \begin{cases} rz\cos\theta & (z \geq 0) \\ rz\sin\theta & (z < 0)\end{cases} \tag{52}$$

so that eq.(51) can be written:

$$(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau})\boldsymbol{\lambda} = J\mathbf{A}^T\mathbf{A}\left[\begin{pmatrix}1\\0\end{pmatrix}\int_0^\infty Dz\, z\,\phi(rz\cos\theta) + \begin{pmatrix}0\\1\end{pmatrix}\int_{-\infty}^0 Dz\, z\,\phi(rz\sin\theta)\right] \tag{53}$$

We now derive further the dynamics for the step transfer function $\phi(x) = R\mathbb{1}_{\{x>0\}}$. Notice that (i) for $z > 0$, $rz\cos\theta > 0 \Leftrightarrow \lambda_1 > 0$; and (ii) for $z < 0$, $rz\sin\theta > 0 \Leftrightarrow \lambda_2 < 0$. The above gives:

$$(\tau\frac{\mathrm{d}}{\mathrm{d}t} + 1)\boldsymbol{\lambda} = \tau JR\langle[z]_+\rangle\mathbf{A}^T\mathbf{A}\left[\begin{pmatrix}1\\0\end{pmatrix}\mathbb{1}_{\{\lambda_1>0\}} - \begin{pmatrix}0\\1\end{pmatrix}\mathbb{1}_{\{\lambda_2<0\}}\right] \tag{54}$$

where $\langle[z]_+\rangle = \int_0^\infty Dz\, z = 1/\sqrt{2\pi}$. Let us define $k = \tau JR\langle[z]_+\rangle$, and:

$$\boldsymbol{\lambda}^{(+)} = k\mathbf{A}^T\mathbf{A}\begin{pmatrix}1\\0\end{pmatrix} = k\begin{pmatrix}1\\ \cos\alpha\end{pmatrix} \quad ; \quad \boldsymbol{\lambda}^{(-)} = -k\mathbf{A}^T\mathbf{A}\begin{pmatrix}0\\1\end{pmatrix} = k\begin{pmatrix}-\cos\alpha\\ -1\end{pmatrix} \tag{55}$$

We can write:

$$\begin{aligned}(\tau\frac{\mathrm{d}}{\mathrm{d}t} + 1)\boldsymbol{\lambda} &= \boldsymbol{\lambda}^{(+)}\mathbb{1}_{\{\lambda_1>0\}} + \boldsymbol{\lambda}^{(-)}\mathbb{1}_{\{\lambda_2<0\}} \\ &= \begin{cases}\boldsymbol{\lambda}^{(+)} & (\boldsymbol{\lambda}\in Q_1)\\ 0 & (\boldsymbol{\lambda}\in Q_2)\\ \boldsymbol{\lambda}^{(-)} & (\boldsymbol{\lambda}\in Q_3)\\ \boldsymbol{\lambda}^{(+)} + \boldsymbol{\lambda}^{(-)} & (\boldsymbol{\lambda}\in Q_4)\end{cases}\end{aligned} \tag{56}$$

with $Q_{1\leq i\leq 4}$ the four quadrants of $\mathbb{R}^2$, numbered in trigonometric order. For $\alpha \in]0, \pi/2[$, $\boldsymbol{\lambda}^{(+)} \in Q_1$, $\boldsymbol{\lambda}^{(-)} \in Q_3$ and $(\boldsymbol{\lambda}^{(+)} + \boldsymbol{\lambda}^{(-)}) \in Q_4$. They are thus stable fixed points of the dynamics, while $\boldsymbol{\lambda} = 0$

is a saddle point. The dynamics of the original variables $\boldsymbol{\kappa}$, shown in fig.3(B), are obtained through the change of coordinates $\boldsymbol{\kappa} = (\mathbf{A}^T)^{-1}\boldsymbol{\lambda}$. In particular, the three corresponding fixed points are:

$$\boldsymbol{\kappa}^{(+)} = k\begin{pmatrix}1\\0\end{pmatrix} \quad ; \quad \boldsymbol{\kappa}^{(-)} = k\begin{pmatrix}-\cos\alpha\\+\sin\alpha\end{pmatrix} \approx k\begin{pmatrix}-1/2\\0.87\end{pmatrix} \tag{57}$$

and $(\boldsymbol{\kappa}^{(+)} + \boldsymbol{\kappa}^{(-)})$. The numerical values correspond to Fig.3(B), where $\alpha = \pi/3$.

B.2 Limit cycles in two different models

Ring model.

For the ring model, we have from eq.(9):

$$\boldsymbol{f}(z) = \begin{pmatrix}\cos z\\\sin z\end{pmatrix} \quad ; \quad \boldsymbol{g} = J\boldsymbol{f}(z+\delta) \tag{58}$$

and the collective dynamics resulting from eq.(21) write:

$$(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau})\boldsymbol{\kappa} = \int_0^{2\pi} \frac{dz}{2\pi}\boldsymbol{g}(z)\phi(\boldsymbol{f}(z)\cdot\boldsymbol{\kappa}) \tag{59}$$

Let ϱ, θ be the polar coordinates of $\boldsymbol{\kappa}$. In the polar coordinate system $(\hat{e}_\varrho, \hat{e}_\theta)$ associated with $\boldsymbol{\kappa}$, we have:

$$(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau})\boldsymbol{\kappa} = (\dot{\varrho} + \frac{1}{\tau}\varrho)\hat{e}_\varrho + \varrho\dot{\theta}\hat{e}_\theta \quad ; \quad \boldsymbol{f}(z) = \cos(z-\theta)\hat{e}_\varrho + \sin(z-\theta)\hat{e}_\theta \tag{60}$$

hence:

$$\begin{aligned}\begin{pmatrix}\dot{\varrho} + \frac{1}{\tau}\varrho\\\varrho\dot{\theta}\end{pmatrix} &= J\int_0^{2\pi}\frac{dz}{2\pi}\begin{pmatrix}\cos(z+\delta-\theta)\\\sin(z+\delta-\theta)\end{pmatrix}\phi(\varrho\cos(z-\theta))\\ &= J\int_0^{2\pi}\frac{dz}{2\pi}\begin{pmatrix}\cos(z+\delta)\\\sin(z+\delta)\end{pmatrix}\phi(\varrho\cos z)\end{aligned} \tag{61}$$

For the step transfer function $\phi(x) = R\mathbb{1}_{\{x>0\}}$, we have $\phi(\varrho\cos z) = R\mathbb{1}_{\{-\pi/2<z<\pi/2\}}$, hence:

$$\begin{pmatrix}\dot{\varrho} + \frac{1}{\tau}\varrho\\\varrho\dot{\theta}\end{pmatrix} = \frac{JR}{2\pi}\int_{-\pi/2}^{\pi/2} dz\begin{pmatrix}\cos(z+\delta)\\\sin(z+\delta)\end{pmatrix} = \frac{JR}{\pi}\begin{pmatrix}\cos\delta\\\sin\delta\end{pmatrix} \tag{62}$$

which is identical to eq.(26).

Gaussian positions.

Let the neurons being normally distributed in $\mathbb{R}^2$, $z_i =\sim \mathcal{N}(0, \mathbb{1}_2)$. As discussed in the Results section 2, the distribution can be parameterised in polar coordinates: each neuron is assigned a radius $\xi_i \sim \chi_2$ and an angle $\varphi_i \sim \mathcal{U}([0, 2\pi[)$. The connectivity of eq.(10) thus writes:

$$W_{ij} = \tilde{J}z_i R^{(\delta)} z_j = \tilde{J}\begin{pmatrix}\xi_i\cos\varphi_i\\\xi_i\sin\varphi_i\end{pmatrix}\begin{pmatrix}\xi_j\cos(\varphi_j+\delta)\\\xi_j\sin(\varphi_j+\delta)\end{pmatrix} \tag{63}$$

which gives the feature maps, as functions of a neuron's radius and angle:

$$\tilde{\boldsymbol{f}}(\xi,\varphi) = \begin{pmatrix}\xi\cos\varphi\\ \xi\sin\varphi\end{pmatrix} = \xi\boldsymbol{f}(\varphi) \quad ; \quad \tilde{\boldsymbol{g}}(\xi,\varphi) = \tilde{J}\tilde{\boldsymbol{f}}(\xi,\varphi+\delta) = \xi\frac{\tilde{J}}{J}\boldsymbol{g}(\varphi) \tag{64}$$

where $\boldsymbol{f},\boldsymbol{g}$ are those of eq.(58). The collective dynamics of eq.(21) thus write:

$$(\frac{\mathrm{d}}{\mathrm{d}t}+\frac{1}{\tau})\boldsymbol{\kappa} = \frac{\tilde{J}}{J}\int_0^{2\pi}\frac{d\varphi}{2\pi}\int_{\mathbb{R}^+} D\xi\ \xi\boldsymbol{g}(\varphi)\phi(\xi\boldsymbol{f}(\varphi)\cdot\boldsymbol{\kappa}) \tag{65}$$

where $D\xi$ denotes the χ_2-probability measure. Notice that for the step transfer function, $\phi(\xi x) = \phi(x)\ \forall\xi>0$, hence:

$$(\frac{\mathrm{d}}{\mathrm{d}t}+\frac{1}{\tau})\boldsymbol{\kappa} = \frac{\tilde{J}}{J}\Big(\int_{\mathbb{R}^+}\xi D\xi\Big)\int_0^{2\pi}\frac{d\varphi}{2\pi}\ \boldsymbol{g}(\varphi)\phi(\boldsymbol{f}(\varphi)\cdot\boldsymbol{\kappa}) \tag{66}$$

which is identical to eq.(59), up to a factor $\frac{\tilde{J}}{J}\langle\xi\rangle$ on the RHS, where $\langle\xi\rangle := \int\xi D\xi$ ($=\sqrt{\pi/2}$ for the χ_2-distribution). The rescaling factor equals 1 if $J=\langle\xi\rangle\tilde{J}$; hence, the two models' collective dynamics are identical up to a rescaling of the connectivity.

Notice that the derivation is not restricted to a normal distribution over $\mathbb{R}^2$, but applies to any rotationally invariant distribution; in this case, the only difference will be the distribution of the radii ξ_i, which will affect the rescaling factor $\langle\xi\rangle$.

General transfer function

Consider the collective dynamics (Eq.(21)) for a general transfer function, corresponding to the connectivity functions of Eq.(64):

$$(\frac{\mathrm{d}}{\mathrm{d}t}+\frac{1}{\tau}\boldsymbol{\kappa}) = \tilde{J}\int_0^{2\pi}\frac{d\varphi}{2\pi}\int_{\mathbb{R}^+} D\xi\ \xi\boldsymbol{f}(\varphi+\delta)\phi(\xi\boldsymbol{f}(\varphi)\cdot\boldsymbol{\kappa}) \tag{67}$$

For the radial coordinate (ϱ,θ) of $\boldsymbol{\kappa}$, it writes (using Eq.(60)):

$$\begin{aligned}\begin{pmatrix}\dot{\varrho}+\frac{1}{\tau}\varrho\\ \varrho\dot{\theta}\end{pmatrix} &= \tilde{J}\int_0^{2\pi}\frac{d\varphi}{2\pi}\int_{\mathbb{R}^+} D\xi\ \xi\begin{pmatrix}\cos(\varphi+\delta-\theta)\\ \sin(\varphi+\delta-\theta)\end{pmatrix}\phi(\xi\varrho\cos(\varphi-\theta))\\ &= \tilde{J}\int_0^{2\pi}\frac{d\varphi}{2\pi}\int_{\mathbb{R}^+} D\xi\ \xi\begin{pmatrix}\cos(\varphi+\delta)\\ \sin(\varphi+\delta)\end{pmatrix}\phi(\xi\varrho\cos(\varphi))\\ &= \tilde{J}\begin{pmatrix}\cos\delta\\ \sin\delta\end{pmatrix}\cdot\Phi(\varrho)\end{aligned} \tag{68}$$

where $\Phi(\varrho)$ is given by:

$$\Phi(\varrho) := \int_{-1}^{1}\frac{dx}{\pi\sqrt{1-x^2}}\int_{\mathbb{R}^+} D\xi\ \xi x\,\phi(\varrho\xi x) \tag{69}$$

Thus, the model generates rotational dynamics that converge towards a limit cycle of radius ϱ_0, implicitly defined by $\varrho_0 = \tau\tilde{J}\cos\delta\Phi(\varrho_0)$ (provided that a solution exists).

Radius parameter as a heterogeneity

The radius ξ_i of each neuron can be interpreted as a heterogeneity, which affects both the connectivity and the transfer functions of the neurons. The connectivity of Eq. (10) reads $W_{ij} = \xi_i \xi_j w^{\text{ring}}(\varphi_i, \varphi_j)$, where w^{ring} denotes the connectivity kernel of the ring model (Eq.(9)). On the other hand, given the polar coordinates (ϱ, θ) of the latent variable $\boldsymbol{\kappa}$, each neuron's firing rate is given by $r_i(t) = \phi(\xi_i \cdot v_t^{\text{ring}}(\varphi_i))$, where $v_t^{\text{ring}}(\varphi_i) = \varrho \cos(\varphi_i - \theta)$ is the potential of a neuron at angle φ_i in the initial ring model.

C Symmetries of the dynamics

We present here the results of Sec.2, relating the symmetries of the field model to those of the collective dynamics. We take an operator theory perspective. The dynamics of the field, eq.(18), can be written:

$$\Box v_t := \left(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau}\right) v_t = \mathcal{D}^{\mathcal{H}} v_t \tag{70}$$

where $\Box = \left(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau}\right)$ is linear, and $\mathcal{D}^{\mathcal{H}} : \mathcal{H} \to \mathcal{F}$ (where $\mathcal{F} = \text{span}\{f_\mu\}$) is a nonlinear operator defined by:

$$\mathcal{D}^{\mathcal{H}} : \varphi \mapsto \sum_\mu f_\mu \langle g_\mu, \phi \circ \varphi \rangle_\rho \tag{71}$$

Symmetries.

Let $S : \mathcal{M} \to \mathcal{M}$ be a bijection such that $\varphi \circ S \in \mathcal{H}$, $\forall \varphi \in \mathcal{H}$. For such a mapping S, $L^{(S)} : \varphi \mapsto \varphi \circ S$ defines a linear operator on $\mathcal{H}$. We call S a *symmetry* of a field model if the operator $\mathcal{D}^{\mathcal{H}}$ describing the dynamics commutes with $L^{(S)}$; i.e. $\mathcal{D}^{\mathcal{H}} \circ L^{(S)} = L^{(S)} \circ \mathcal{D}^{\mathcal{H}}$. It implies the equivariance of the field dynamics for $L^{(S)}$: if v_t solves the dynamics of eq.(70), then $L^{(S)} v_t$ is also a solution to the dynamics, with initial condition $L^{(S)} v_0$.

Similarly, the collective dynamics of eq.(21) are described by a nonlinear operator $\mathcal{D}^D$ acting on $\mathbb{R}^D$:

$$\begin{aligned} \Box \boldsymbol{\kappa} &= \mathcal{D}^D \boldsymbol{\kappa} \\ \text{with} \quad \mathcal{D}^D &: \boldsymbol{\kappa} \mapsto \langle \boldsymbol{g}, \phi \circ (\boldsymbol{f} \cdot \boldsymbol{\kappa}) \rangle_\rho \end{aligned} \tag{72}$$

and a symmetry of the collective dynamics is a mapping $T : \mathbb{R}^D \to \mathbb{R}^D$ that commutes with $\mathcal{D}^D$.

Theorem 1. *1. If (i) S is a symmetry of the field model, and (ii) $\mathcal{F} = \text{span}\{f_\mu\}$ is an invariant subspace of $L^{(S)}$, then a symmetry of the collective dynamics is given by the linear operator $T^{(S)} : \mathbb{R}^D \to \mathbb{R}^D$, defined by the matrix*

$$T^{(S)} = M(\langle f_\mu, f_\nu \circ S \rangle_\rho) \tag{73}$$

where $M^{-1} = (\langle f_\mu, f_\nu \rangle_\rho)$, and $(a_{\mu\nu})$ denotes the $D \times D$ matrix with entries $a_{\mu\nu}$.

2. If both the distribution ρ and the connectivity kernel $w(z, z') = \boldsymbol{f}(z) \cdot \boldsymbol{g}(z')$ are invariant for S, then the requirements (i) and (ii) of point 1 are satisfied.

The proof can be found at the end of this section.

Left and right symmetry groups.

Let $S : (g, z) \mapsto S_g(z)$ denote the action of a group G on $\mathcal{M}$. G is a *symmetry group* of the field model if for any $g \in G$, S_g is a symmetry. We call it a *left* symmetry group if S is a left group action (for which $S_{gh} = S_g \circ S_h$, i.e. $S_{gh}(z) = S_g(S_h(z))$, $\forall g, h \in G$); and a *right* symmetry group for a

right group action (for which $S_{gh} = S_h \circ S_g$). Similarly, a symmetry group of the collective dynamics is a group acting on $\mathbb{R}^D$ whose action of any element is a symmetry of the collective dynamics.

Corollary 1. *Let G be a* left *symmetry group of the field model, whose action S_g of each element g on $\mathcal{M}$ satisfies the requirements of Theorem 1.1. Then, G is a* right *symmetry group of the collective dynamics, with a linear group action on $\mathbb{R}^D$ given by $(g, \boldsymbol{\kappa}) \mapsto T^{(S_g)}\boldsymbol{\kappa}$.*

Proofs.

In the following, $S : \mathcal{M} \to \mathcal{M}$ denotes a bijection satisfying $\varphi \circ S \in \mathcal{H}$, $\forall \varphi \in \mathcal{H}$; so that $L^{(S)} : \mathcal{H} \to \mathcal{H} : \varphi \mapsto \varphi \circ S$ is well defined.

We start by defining the surjection $K : \mathcal{H} \to \mathbb{R}^D$ that maps a field to the components of its projection on $\mathcal{F}$, and the reciprocal injection $V : \mathbb{R}^D \to \mathcal{F} \subset \mathcal{H}$:

$$K : \varphi \mapsto M\langle \boldsymbol{f}, \varphi\rangle_\rho \quad ; \quad V : \boldsymbol{k} \mapsto \boldsymbol{k} \cdot \boldsymbol{f} \tag{74}$$

with $M^{-1} = (\langle f_\mu, f_\nu\rangle_\rho)$ the inverse correlation matrix. They satisfy the following properties:

$$KV = \mathbb{1}_{\mathbb{R}^D} \quad ; \quad VK = \Pi_{\mathcal{F}} \tag{75}$$

where $\Pi_{\mathcal{F}} : \varphi \mapsto \sum_{\mu,\nu} M_{\mu\nu} f_\mu \langle f_\nu, \varphi\rangle_\rho$ is the orthogonal projection on $\mathcal{F}$ (by definition of the matrix M). Thus, $K : v \mapsto \boldsymbol{\kappa}$ and $V : \boldsymbol{\kappa} \mapsto v$ establish a bijection between $\mathcal{F}$ and $\mathbb{R}^D$.

We start with a few useful lemmas.

Lemma 1. $\mathcal{D}^D = K\mathcal{D}^{\mathcal{H}}V$.

Proof. By setting $\boldsymbol{\kappa}(t) = Kv_t$ and $v_t = V\boldsymbol{\kappa}(t)$, and noticing that $\Box = (\partial_t - 1/\tau)$ commutes with K, one obtains from eq.(70) the following expression for the dynamics of $\boldsymbol{\kappa}$:

$$\Box\boldsymbol{\kappa} = \Box K v_t = K\Box v_t = K\mathcal{D}^{\mathcal{H}} v_t = K\mathcal{D}^{\mathcal{H}} V\boldsymbol{\kappa}(t) \tag{76}$$

which can be identified with eq.(72)). □

Lemma 2. *If $\mathcal{F}$ is an invariant subspace of an operator L on $\mathcal{H}$, then $LV = \Pi_{\mathcal{F}} LV$.*

Proof. For any $\varphi \in \mathcal{H}, V\varphi \in \mathcal{F}$. $\mathcal{F}$ is an invariant subspace of L iff $Lv \in \mathcal{F}$ $\forall v \in \mathcal{F}$. Then $\forall \varphi \in \mathcal{H}$, $LV\varphi \in \mathcal{F}$, implying $\Pi_{\mathcal{F}} LV\varphi = LV\varphi$. □

Lemma 3. *If ρ is invariant for the bijection S, then $(L^{(S)})^\dagger = L^{(S^{-1})}$.*

Proof. ${L^{(S)}}^\dagger$ is defined by: $\langle \varphi, L^{(S)}\psi\rangle_\rho = \langle {L^{(S)}}^\dagger \varphi, \psi\rangle_\rho$ for all $\varphi, \psi \in \mathcal{H}$. If ρ is invariant for S, i.e. $\rho \circ S^{-1} = \rho$, then for all $\varphi, \psi \in \mathcal{H}$,

$$\begin{aligned}
\langle \varphi, L^{(S)}\psi\rangle_\rho &= \langle \varphi, \psi \circ S\rangle_\rho = \int_{\mathcal{M}} \varphi(z)\psi(Sz)\rho(\mathrm{d}z) \\
&= \int_{\mathcal{M}} \varphi(S^{-1}z)\psi(z)\,(\rho \circ S^{-1})(dz) = \int_{\mathcal{M}} \varphi(S^{-1}z)\psi(z)\,\rho(\mathrm{d}z) \\
&= \langle \varphi \circ S^{-1}, \psi\rangle_\rho = \langle L^{(S^{-1})}\varphi, \psi\rangle_\rho
\end{aligned} \tag{77}$$

where we used the change of variable formula in the third equality. □

Proof of Theorem 1.

Point 1.

Let L be a linear operator on $\mathcal{H}$ that satisfies: (i) $\mathcal{D}^{\mathcal{H}}L = L\mathcal{D}^{\mathcal{H}}$, and (ii) $\mathcal{F}$ is an invariant subspace of L. We have from (i):

$$K(L\mathcal{D}^{\mathcal{H}})V = K(\mathcal{D}^{\mathcal{H}}L)V \tag{78}$$

The left-hand side can be written:

$$KL\mathcal{D}^{\mathcal{H}}V = KL(VK\mathcal{D}^{\mathcal{H}})V = (KLV)(K\mathcal{D}^{\mathcal{H}}V) = T\mathcal{D}^{D} \tag{79}$$

with $T := KLV$. The first equality is obtained by noticing that the image of $\mathcal{D}^{\mathcal{H}}$ is in $\mathcal{F}$, so that $\Pi_{\mathcal{F}}\mathcal{D}^{\mathcal{H}} = \mathcal{D}^{\mathcal{H}}$; combined with $VK = \Pi_{\mathcal{F}}$, eq.(75); and we used Lemma 1 in the last equality.

In turn, the right-hand side writes:

$$K\mathcal{D}^{\mathcal{H}}(LV) = K\mathcal{D}^{\mathcal{H}}(VKLV) = (K\mathcal{D}^{\mathcal{H}}V)(KLV) = \mathcal{D}^{D}T \tag{80}$$

where we used Lemma 2, with $VK = \Pi_{\mathcal{F}}$.

The operator $T = KLV$ thus commutes with the operator $\mathcal{D}^{D}$: it is a symmetry of the collective dynamics. Given a transformation S such that $L = L^{(S)}$, we can write, for all $\boldsymbol{\kappa} \in \mathbb{R}^{D}$:

$$\begin{aligned}(T\boldsymbol{\kappa})_{\mu} &= \sum_{\nu} M_{\mu\nu}\langle f_{\nu}, L^{(S)}(\sum_{\sigma} f_{\sigma}\kappa_{\sigma})\rangle_{\rho} = \sum_{\nu} M_{\mu\nu}\langle f_{\nu}, \sum_{\sigma}\kappa_{\sigma}(L^{(S)}f_{\sigma})\rangle_{\rho} \\ &= \sum_{\sigma}\sum_{\nu} M_{\mu\nu}\langle f_{\nu}, f_{\sigma}\circ S\rangle_{\rho}\kappa_{\sigma} =: \sum_{\sigma} T^{(S)}_{\mu\sigma}\,\kappa_{\sigma}\end{aligned} \tag{81}$$

which yields $T^{(S)}_{\mu\sigma} = \sum_{\nu} M_{\mu\nu}\langle f_{\nu}, f_{\sigma}\circ S\rangle_{\rho}$. We used the linearity of $L^{(S)}$ in the second equality.

Point 2.

Let both ρ and the connectivity w be invariant for S (i.e. $w(z,z') = w(Sz,Sz')\ \forall z,z' \in \mathcal{M}$). We first show that it implies: $\mathcal{D}^{\mathcal{H}}L^{(S)} = L^{(S)}\mathcal{D}^{\mathcal{H}}$. For all $v \in \mathcal{H}$ and $z \in \mathcal{M}$, we have:

$$\begin{aligned}(\mathcal{D}^{\mathcal{H}}L^{(S)}v)(z) &= \sum_{\mu} f_{\mu}(z)\langle g_{\mu}, \phi\circ(L^{(S)}v)\rangle_{\rho} \\ &= \sum_{\mu} f_{\mu}(z)\langle g_{\mu}, \phi\circ v\circ S\rangle_{\rho} = \sum_{\mu} f_{\mu}(z)\langle g_{\mu}, L^{(S)}(\phi\circ v)\rangle_{\rho} \\ &= \sum_{\mu} f_{\mu}(z)\langle {L^{(S)}}^{\dagger}g_{\mu}, \phi\circ v\rangle_{\rho} \overset{(*)}{=} \sum_{\mu} f_{\mu}(z)\langle g_{\mu}\circ S^{-1}, \phi\circ v\rangle_{\rho} \\ &= \int_{\mathcal{M}} w(z, S^{-1}z')\phi\circ v(z)\rho(\mathrm{d}z) \overset{(**)}{=} \int_{\mathcal{M}} w(Sz,z')\phi\circ v(z)\rho(\mathrm{d}z) \\ &= \sum_{\mu} f_{\mu}(Sz)\langle g_{\mu}, \phi\circ v\rangle_{\rho} = \sum_{\mu}(L^{(S)}f_{\mu})(z)\langle g_{\mu}, \phi\circ v\rangle_{\rho} \\ &= (L^{(S)}\mathcal{D}^{\mathcal{H}}v)(z)\end{aligned} \tag{82}$$

where we used Lemma 3 in $(*)$, and the invariance of w in $(**)$.

We now show that it implies that $\mathcal{F}$ is invariant under $L^{(S)}$. For this, it is sufficient to show that the images of the basis functions $L^{(S)} f_\mu \in \mathcal{F},\ \forall \mu = 1, \dots, D$. We have:

$$\begin{aligned} w \text{ invariant for } S \ &\Leftrightarrow\ \sum_\mu f_\mu(Sz) g_\mu(Sz') = \sum_\mu f_\mu(z) g_\mu(z') \quad \forall z, z' \in \mathcal{M} \\ &\Leftrightarrow\ \sum_\mu (L^{(S)} f_\mu)(z)(L^{(S)} g_\mu)(z') = \sum_\mu f_\mu(z) g_\mu(z') \quad \forall z, z' \\ &\overset{(*)}{\Rightarrow}\ \sum_\mu (L^{(S)} f_\mu)(z) \langle L^{(S)} g_\nu, L^{(S)} g_\mu \rangle_\rho = \sum_\mu f_\mu(z) \langle L^{(S)} g_\nu, g_\mu \rangle_\rho \quad \forall z, \forall \nu \\ &\Rightarrow\ \sum_\mu G_{\nu\mu} L^{(S)} f_\mu = \sum_\mu f_\mu \langle L^{(S)} g_\nu, g_\mu \rangle_\rho \in \mathcal{F} \quad \forall \nu = 1, \dots, D \end{aligned} \tag{83}$$

where $(*)$ comes by projection of the functions of z' on $L^{(S)} g_\nu$; and $G_{\nu\mu} := \langle L^{(S)} g_\nu, L^{(S)} g_\mu \rangle_\rho = \langle g_\nu, g_\mu \rangle_\rho$ from the invariance of ρ and Lemma 3.

Without loss of generality, one can assume that the functions $\{g_\mu\}$ are linearly independent. If not verified, one can always choose a new basis of $\mathcal{F}$ such that this is the case. Thus, the matrix G has full rank, and is therefore invertible. The last line of eq.(83) yields:

$$L^{(S)} f_\mu = \sum_\nu (G^{-1})_{\mu\nu} \sum_\sigma \langle L^{(S)} g_\nu, g_\sigma \rangle_\rho \, f_\sigma \in \mathcal{F} \quad \forall \mu = 1, \dots, D \tag{84}$$

i.e. $L^{(S)} f_\mu \in \mathcal{F},\ \forall \mu = 1, \dots, D$, which concludes the proof.

Proof of Corollary 1.

In the following, we use the shorthand notation $L^{(S_g)} = L^{(g)}$ and $T^{(S_g)} = T^{(g)}$. To prove the corollary, it is sufficient to show that $T^{(gh)} = T^{(h)} T^{(g)}$, $\forall g, h \in G$; and that $T^{(e)} = \mathbb{1}_{\mathbb{R}^D}$ for the identity $e \in G$.

Given that S is a left group action, we have $S_{gh} = S_g \circ S_h$; and thus, $L^{(gh)} = L^{(S_{gh})} = L^{(S_g \circ S_h)}$. We have by associativity of the composition: $L^{(S_g \circ S_h)} \varphi = \varphi \circ (S_g \circ S_h) = (\varphi \circ S_g) \circ S_h = L^{(h)}(L^{(g)} \varphi)$, for all $\varphi \in \mathcal{H}$; hence: $L^{(gh)} = L^{(h)} L^{(g)}$. Thus, $(g, \varphi) \mapsto L^{(g)} \varphi$ defines a *right* group action of G on $\mathcal{H}$.

By recalling $T^{(S)} = K L^{(S)} V$, we get:

$$T^{(gh)} = K L^{(gh)} V = K L^{(h)} L^{(g)} V = K L^{(h)} (VK) L^{(g)} V = T^{(h)} T^{(g)} \tag{85}$$

where we used Lemma 2 in the third equality, with $VK = \Pi_{\mathcal{F}}$ (eq(75)).

Notice eventually that the identity e yields identity group action: we have $S_e = \mathbb{1}_{\mathcal{M}} \ \Rightarrow\ L^{(e)} = \mathbb{1}_{\mathcal{H}} \ \Rightarrow\ T^{(e)} = K \mathbb{1}_{\mathcal{H}} V = \mathbb{1}_{\mathbb{R}^D}$ according to eq.(75).

D Adding heterogeneity in a field model

Rescaling of outgoing weights

Consider a field model, obeying the low-dimensional dynamics of Eq.(21), in which the outgoing weights of each neuron i are rescaled by an heterogeneity parameter $\xi_i > 0$, drawn i.i.d. across

neurons according to a distribution ρ_ξ. In the resulting "new" field model, each neuron's position in the "new" functional similarity space now contains this heterogeneity parameter; we denote it

$$\tilde{z}_i = (z_i, \xi_i) \in \mathcal{M} \times \mathbb{R}_+ =: \tilde{\mathcal{M}} \tag{86}$$

and the "new" connectivity reads:

$$\tilde{W}_{ij} = \tilde{\boldsymbol{f}}(\tilde{z}_i) \cdot \tilde{\boldsymbol{g}}(\tilde{z}_j) = \xi_j W_{ij} = \boldsymbol{f}(z_i) \cdot \xi_j \boldsymbol{g}(z_j) \tag{87}$$

which allows us to identify $\tilde{\boldsymbol{f}}(\tilde{z}_i) = \boldsymbol{f}(z_i)$, and $\tilde{\boldsymbol{g}}(\tilde{z}_j) = \xi_j \boldsymbol{g}(z_j)$. The collective dynamics, Eq.(21), of the latent variables $\tilde{\boldsymbol{\kappa}}$ associated with this "new" model with rescaled connectivity write:

$$(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau})\tilde{\boldsymbol{\kappa}} = \Big(\int \xi \rho_\xi(d\xi) \Big) \cdot \langle \boldsymbol{g}, \phi \circ (\boldsymbol{f} \cdot \tilde{\boldsymbol{\kappa}}) \rangle_\rho \tag{88}$$

which is identical to the dynamics of the original model, up to a rescaling of the latent variables by a factor $\int \xi \rho_\xi(d\xi)$, which represents the mean value of the heterogeneity parameter ξ.

Heterogeneous maximum firing rates

In a similar fashion, starting from an original field model, each neuron's firing rate can be rescaled by a parameter $a_i > 0$ drawn according to a distribution ρ_a. The firing rate of any neuron i, given its membrane potential h_i, is now given by $r_i = a_i \phi(h_i)$. The transfer function $\tilde{\phi}$ of the "new" model thus depends on each neuron's "new" position $\tilde{z}_i = (z_i, a_i)$: we write $\tilde{\phi}(h, \tilde{z}_i) = a_i \phi(h)$. Similar to the example above ("Rescaling of output weights"), the collective dynamics of the "new" and of the original field models are identical, up to a rescaling by the mean value of the heterogeneity parameter $\int a \rho_a(da)$.

E Field models as universal approximators

E.1 Gaussian mixture low-rank networks

Gaussian mixture low-rank networks are a class of low-rank RNNs recently introduced to study how the collective dynamics are shaped by the statistics of the connectivity vectors [41, 54]. In this model, the rank-D interaction between neurons i, j is given by $\boldsymbol{m}_i \cdot \boldsymbol{n}_j$, where the connectivity vectors $\boldsymbol{m}_i, \boldsymbol{n}_i \in \mathbb{R}^D$ associated with each neuron i are drawn according to the Gaussian mixture of P populations:

$$z_i = (\boldsymbol{m}_i, \boldsymbol{n}_i) \sim \rho = \sum_{p=1}^{P} \alpha_p \rho^{(p)} \tag{89}$$

where for each $p = 1 \ldots, P$, $\rho^{(p)}$ is a multivariate Gaussian over $\mathbb{R}^{2D}$; and $\sum_{p=1}^{P} \alpha_p = 1$. Here, each neuron's position $z_i \in \mathbb{R}^{2D}$ corresponds to the concatenation of the connectivity vectors, and one can directly identify $\boldsymbol{f}(\boldsymbol{m}, \boldsymbol{n}) = \boldsymbol{m}$, $\boldsymbol{g}(\boldsymbol{m}, \boldsymbol{n}) = \boldsymbol{n}$.

These RNNs have been shown to be universal approximators of arbitrary D-dimensional collective dynamics, if the number P of populations can be chosen arbitrarily [54].

E.2 Field models over one-dimensional space

For the sake of simplicity, we focus on the case of a bounded one-dimensional space $\mathcal{M} = [0, 1]$, with uniform distribution $\rho(\mathrm{d}z) = dz$. We furthermore restrict to the subspace of the connectivity functions that vanish at the boundary, that we denote $\mathcal{H}_0$.

First of all, biases can be introduced by adding a constant latent variable $\kappa_0(t) = 1$, associated with a bias function $f_0(z)$. Let the set of approximating functions, called AF_0, be the set of functions $\hat{D} : \mathbb{R}^D \to \mathbb{R}$ of the form

$$\hat{D}(\boldsymbol{\kappa}) = \int_0^1 g(z)\phi\Big(\sum_{\nu=1}^{D} \kappa_\nu f_\nu(z) + f_0(z)\Big)dz \qquad \text{with } g, \{f_\mu\}_{0\leq\mu\leq D} \subset \mathcal{H}_0 \tag{90}$$

AF_0 is a subset of the linear readouts that can be generated by field models over bounded one-dimensional space.

Consider an arbitrary dynamical system with asymptotic exponential decay:

$$\frac{\mathrm{d}}{\mathrm{d}t}\kappa_\alpha(t) = -\frac{1}{\tau}\kappa_\alpha(t) + D_\alpha(\vec{\kappa}(t)) \quad (\alpha = 1, \dots, D)$$

where $D_\alpha \in \mathcal{C}(\mathbb{R}^D, \mathbb{R})$ vanish for $\|\vec{\kappa}\| \to \infty$. The approximation problem consists in finding functions g and $\{f_\mu\}$ so that each given D_μ is approximated by some $\hat{D} \in \mathrm{AF}_0$. As for FFNNs, it can be decomposed into two parts [87]. First, show that approximating functions can be linearly combined: for any $\hat{D}^{(1)}, \hat{D}^{(2)} \in \mathrm{AF}_0$ and $c_1, c_2 \in \mathbb{R}$, one can find $\hat{D}' \in \mathrm{AF}_0$ such that $\hat{D}' = c_1\hat{D}^{(1)} + c_2\hat{D}^{(2)}$. Second, show that elementary functions (like half-plane indicators) can be well approximated. Arbitrary functions can then be approximated using linear combinations of elementary functions.

Linear combination.

For any linear combination in AF_0, we can write:

$$\begin{aligned}
(c_1\hat{D}^{(1)} + c_2\hat{D}^{(2)})(\vec{\kappa}) &= c_1 \int_0^1 g^{(1)}(z)\phi\left(\vec{\kappa}\cdot\vec{f}^{(1)}(z) + f_0^{(1)}(z)\right)dz \\
&\quad + c_2 \int_0^1 g^{(2)}(z)\phi\left(\vec{\kappa}\cdot\vec{f}^{(2)}(z) + f_0^{(2)}(z)\right)dz \\
&= c_1 \int_0^{1/2} 2g^{(1)}(2z)\phi\left(\vec{\kappa}\cdot\vec{f}^{(1)}(2z) + f_0^{(1)}(2z)\right)dz \\
&\quad + c_2 \int_{1/2}^1 2g^{(2)}(2z-1)\phi\left(\vec{\kappa}\cdot\vec{f}^{(2)}(2z-1) + f_0^{(2)}(2z-1)\right)dz \\
&= \int_0^1 g'(z)\phi\left(\vec{\kappa}\cdot\vec{f}'(z) + f_0'(z)\right)dz \quad =: \hat{D}'(\vec{\kappa})
\end{aligned}$$

where the parameters of $\hat{D}'$ are

$$g'(z) = \begin{cases} 2c_1 g^{(1)}(2z) & \text{if } z \in [0, 1/2[\\ 2c_2 g^{(2)}(2z-1) & \text{if } z \in [1/2, 1] \end{cases} \; ; \; e'_\alpha(z) = \begin{cases} e_\alpha^{(1)}(2z) & \text{if } z \in [0, 1/2[\\ e_\alpha^{(2)}(2z-1) & \text{if } z \in [1/2, 1] \end{cases}$$

for $\alpha = 0, \ldots, D$. Because of the vanishing boundary condition on $[0,1]$, if $g^{(i)}, \{f_\mu^{(i)}\}$ are in $\mathcal{H}_0$, so are $g', \{f'_\mu\}$, and so $\hat{D}' \in \mathrm{AF}_0$. Thus, AF_0 is a vector subspace of $\mathcal{C}(\mathbb{R}^D, \mathbb{R})$.

Elementary functions.

Following the line of [87], we can build approximating functions for the indicator of half-planes in $\mathbb{R}^D$, i.e. $\mathbb{1}\{\boldsymbol{\kappa} \cdot \boldsymbol{d} > b\}$ for arbitrary $\boldsymbol{d} \in \mathbb{R}^D, b \in \mathbb{R}$.

Let $\mathcal{H}_0 \ni T : z \mapsto 2(1 - |2z - 1|)$ be the normalised triangle function over $[0,1]$. Given $\boldsymbol{d} \in \mathbb{R}^D$ and $b \in \mathbb{R}$, for $A > 0$, let $\boldsymbol{e}(z) = AT(z)\boldsymbol{d}$, $e_0(z) = -AT(z)b$ and $g(z) = T(z)$. Then for $A \to \infty$,

$$\hat{D}_A(\boldsymbol{\kappa}) := \int_0^1 T(z)\phi\left(AT(z)\big(\boldsymbol{\kappa} \cdot \boldsymbol{d} - b\big)\right) dz \to \begin{cases} \phi(+\infty) & (\boldsymbol{\kappa} \cdot \boldsymbol{d} > b) \\ \phi(-\infty) & (\boldsymbol{\kappa} \cdot \boldsymbol{d} < b) \\ \phi(0) & (\boldsymbol{\kappa} \cdot \boldsymbol{d} = b) \end{cases}$$

By assuming that ϕ is sigmoidal (i.e. $\phi(+\infty) = 1, \phi(-\infty) = 0$), we see that $\hat{D}_A$ approximates the indicator of the half-plane in direction $\boldsymbol{d}$, given A large enough. The constant function $\hat{D}(\kappa) = 1$ can also be trivially constructed.

Since linear combinations of indicators of half-planes are dense in $C(\mathbb{R}^D, \mathbb{R})$ with compact support [87], arbitrary functions D_α can be approximated by functions of the form of eq.(90).