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 jointly supervised this work.

Abstract

The classic view of cortical circuits composed of precisely tuned neurons hardly accounts for large-scale recordings indicating that neuronal populations are heterogeneous and exhibit activity patterns evolving on low-dimensional manifolds. Using a modelling approach, we connect these two contrasting views. Our recurrent spiking network models explicitly link the circuit structure with the low-dimensional dynamics of the population activity. Importantly, we show that different circuit models can lead to equivalent low-dimensional dynamics. Nevertheless, we design a method for retrieving the circuit structure from large-scale recordings and test it on simulated data. Our approach not only unifies cortical circuit models with established models of collective neuronal dynamics, but also paves the way for identifying elements of circuit structure from large-scale experimental recordings.

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

Introduction

The progress of experimental methods continuously reshapes our understanding of the link between brain activity and behaviour. 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 activity of single neurons can be modulated by features of sensory stimuli [4–6], spatial location 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 termed *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, 17]. Single-neuron properties might also describe other response features besides stimulus tuning: neuronal responses can be modulated by some task- or stimulus-independent variables that are not controlled by the experimenter, such as internal brain state [18–22]. In this view, each neuron can be assigned 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 briefly *similarity space*. Importantly, the embedding of the neurons may not correspond to their physical location in the brain [22]. For instance in mammalian V1, preferred orientations are distributed across the cortical sheet, forming pinwheel-like patterns [23, 24].

The second perspective takes a descriptive, population-level approach, put forward by the advent of large-scale recordings [25]. In this perspective, the joint activity of N simultaneously recorded neurons, or *population activity*, is represented at any given time as a point in a N-dimensional state-space. Several studies have 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 *high-dimensional* state-space of the population activity [26–31]. In the neural manifold, the momentary activity can be described by a small number $D \ll N$ of time-dependent variables that represent computationally relevant quantities; e.g., the animal's head direction, or the perceived sensory stimulus (Fig. 1C). The change of these variables over time reveals the low-dimensional dynamics of the population activity on a time scale from tens of milliseconds [21, 32–34] to tens of seconds [22] (Fig. 1D). Hence, in this view, neural computation is described by a set of trajectories, or "flow", in the low-dimensional neural manifold (Fig. 1E) that emerges from the collective dynamics of a large number of neurons [34–38].

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 provides a direct link between the two perspectives [39]. 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 the 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

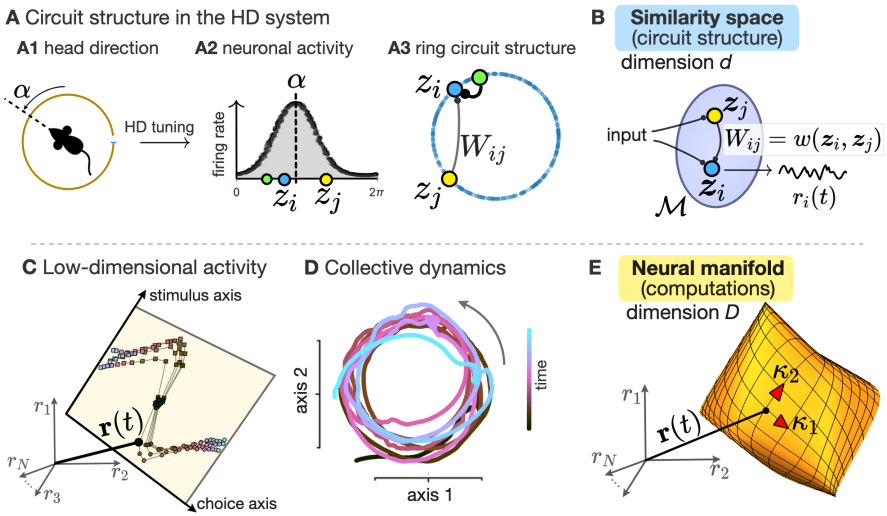

Fig. 1 *Two perspectives on neuronal activity.* **A.** The "circuit" perspective illustrated by the ring model of the head direction (HD) system. **A1.** The animal's HD is represented by an angle α. **A2.** Each neuron i in the population responds preferentially to a specific HD angle z_i (horizontal axis). Neuronal activity (vertical axis) depends on the difference between the current HD α and each neuron's preferred HD angle z_i. Coloured dots represent three examples neurons. **A3.** Each neuron's preferred HD angle determines its "location" on a one-dimensional ring (blue dots). The persistent activity of HD cells in the absence of stimulation [42, 46] can be explained by a ring attractor model [16, 17], where recurrent synaptic connections W_{ij} between pairs of neurons i and j are stronger if neurons are close neighbours on the ring (e.g. the blue and green neuron) than if they are further apart (e.g. the blue and yellow neuron). **B.** In the "circuit" perspective, single-neuron responses $r_i(t)$ are characterised by the locations of the neurons in an abstract *similarity space* $\mathcal{M}$ 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 inputs) are determined by their locations $\boldsymbol{z}_i$, $\boldsymbol{z}_j$ in this abstract space. **C.** The "neural manifold" perspective illustrated by large-scale recordings of the neuronal activity in the monkey prefrontal cortex [34]. During a decision-making task, the population activity $\mathbf{r}(t)$ of a population of N cells lives in a two-dimensional subspace, spanned by a 'stimulus' and a 'choice' axis, which encode information about the sensory stimulus and the decision made during a trial, respectively. Connected points correspond to the population activity during a single trial, with color representing time elapsed since trial onset. **D.** In mouse medial entorhinal cortex, the linear projection of the population activity onto relevant axes reveals slow oscillatory population-level dynamics, internally generated by recurrent mechanisms [22]. **E.** In the "manifold" perspective, the joint activity of N neurons, $\mathbf{r}(t) \in \mathbb{R}^N$, lies on a low-dimensional *neural manifold* with intrinsic dimension $D \ll N$ (shown here 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 (C) is inspired from [20]. Panel (D) is inspired from [22].

variable (for example, the head direction in the case of HD cells), the set of all the possible population activities will in turn necessarily form a ring. More generally, low-dimensional tasks tend to yield low-dimensional neural trajectories [40, 41]. Yet, such a static transformation from stimulus to population activity is insufficient to explain some of the most striking features of neural manifolds. First, it has been observed that neural manifolds can persist during sleep [27, 28, 42], and be dissociated from sensory cues [10, 43]; which implies 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 present external stimulus variables, but is rather determined by the initial condition [21], context [34], or internally-generated factors [22, 30, 44] (Fig. 1D). Both findings point towards underlying recurrent circuit mechanisms [26, 45].

In the examples mentioned above, strong excitatory connections are found between neurons with similar preferred HD in the fly HD system [47], or between neurons with a similar orientation tuning in V1 [24]. Consistent with the traditional notion of cortical columns [48], these observations support the idea that functional similarities between neurons arise from network structure [12, 42, 49]. This idea has been the basis for many computational models in which forward and recurrent connections are determined by the neurons' respective functional properties (Fig. 1B) [50–57]. However, the link between so-called *circuit models* and neural manifolds remains elusive [2, 3].

In this paper, our aim is to connect the two perspectives by addressing the following questions. (i) Can we establish a mathematical link between the similarity space, summarising the structure of the connectivity in circuit models, and the flow of the collective dynamics in the neural manifold? (ii) Is there a one-to-one mapping between the neural manifold and the similarity space? Although the topology of the neural manifold has been found to match that of the similarity space in special cases (for example, a ring in the fly HD system) [3, 47], there is no obvious reason why this should be true in general. (iii) Can we extract knowledge about the circuit structure from neuronal activity? Although many studies have focused on extracting the neural manifold from experimental data [27, 28, 32, 33], 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 modelling framework that provides an explicit link between circuit structure and collective dynamics in Recurrent Neural Networks (RNNs). Relying on the theory of neural fields [52] and large low-rank networks [58], our framework connects functional properties of single neurons to emergent low-dimensional dynamics in an RNN model. 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, and the same collective dynamics can result from different circuit structures. Finally, we propose a method for extracting the similarity space from the neuronal activity, by exploiting the similarities between the activity of single neurons. We apply this method to simulated data, and show that we can retrieve the dissimilar circuit structure of different RNNs performing the same simple task.

Results

A unifying framework based on the theory of large RNNs

In this section, we present a modelling framework that links 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 1. Our framework extends recent theoretical results on large RNNs [59–61] to field models [52, 59]. Since the dynamics of large Spiking Neural Networks (SNNs) composed of Linear-Nonlinear-Poisson (LNP) neurons [62, 63] converges to the dynamics of equally large RNNs in the large N limit (with $1/N$ synaptic weight scaling) [61], we directly consider the case of 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 integrator (with time constant τ) of recurrent and external inputs. Each spike of a presynaptic neuron j, connected to postsynaptic neuron i, causes an instantaneous change of the postsynaptic potential h_i with magnitude given by the synaptic weight $\frac{1}{N}W_{ij}$ 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 $w_{i1}^{\text{ext}}, \ldots, w_{iK}^{\text{ext}}$, respectively. The dynamics of the membrane potential h_i thus 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 $S_j(t) = \sum_{s \in \mathbb{N}} \delta(t - t_j^s)$ denotes the spike train of neuron j. The instantaneous firing rate of each neuron i is given by $r_i(t) = \phi(h_i(t))$, where ϕ is a positive, *nonlinear* transfer function. The set of instantaneous firing rates, $\mathbf{r}(t) = (r_1(t), \ldots, r_N(t))$, define the intensity of N (conditionally independent) inhomogenous *Poisson* processes from which the spike times $\{t_j^s\}$ of each neuron j are drawn.

The theoretical framework relies on two assumptions. First, we assume that each neuron i can be functionally characterised by a small set of $d \ll N$ features, denoted $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,d})$, which determine its abstract location in a d-dimensional *functional similarity space* $\mathcal{M}$, where neurons close to each other have similar response properties. This assumption is rooted in the circuit perspective described in the introduction and is the starting point of many models in computational neuroscience [16, 50–57]. This assumption means that both recurrent and external connections in Eq.(1) depend on the respective "locations" of the neurons, so that Eq.(1) can be rewritten as:

$$\frac{\mathrm{d}}{\mathrm{d}t} h_i(t) = -\frac{1}{\tau} h_i(t) + \frac{1}{N} \sum_{j=1}^{N} w(\boldsymbol{z}_i, \boldsymbol{z}_j) S_j(t) + \sum_{\nu=1}^{K} f_\nu^{\text{ext}}(\boldsymbol{z}_i) I_\nu(t), \tag{2}$$

where $f_1^{\text{ext}}, \ldots, f_K^{\text{ext}}$ are continuous functions that characterise the tuning of the neurons to the external inputs $I_1, \ldots, I_K$ as a function of their functional properties. Similarly, the connectivity kernel w describes the recurrent connectivity as a function of the neurons' functional properties.

Second, we assume that the kernel w can be written in the form:

$$W_{ij} = w(\boldsymbol{z}_i, \boldsymbol{z}_j) = \sum_{\mu=1}^{D} f_\mu(\boldsymbol{z}_i) g_\mu(\boldsymbol{z}_j). \tag{3}$$

where $f_1, \ldots, f_D$ and $g_1, \ldots, g_D$ are arbitrary continuous functions over the similarity space $\mathcal{M}$ (Fig. 2); and $D \ll N$ is a fixed number, not necessarily related to the dimension d of the similarity space. Motivated by results on low-rank RNNs, this assumption is rooted in the "neural manifold" perspective: in an RNN where the recurrent connectivity has rank D, the population activity lies in a D-dimensional neural manifold [58]. Models of low-rank networks can perform a wide range of computations, from the retrieval of patterns stored as fixed points (as in the Hopfield model) [64–66] to the approximation of arbitrary dynamical systems [38, 60]. We note that the matrix of recurrent connections in Eq.(3) can be written as the outer product of two $N \times D$ matrices F and G:

$$W = F \cdot G^T \tag{4}$$

with entries $F_{i,\mu} = f_\mu(\boldsymbol{z}_i)$ and $G_{j,\mu} = g_\mu(\boldsymbol{z}_j)$. Therefore, we recognise the well-known structure of a connectivity with rank D (Fig. 2) [58].

Existing models of large low-rank networks assume that the rows $\{F_{i,\mu}\}, \{G_{i,\mu}\}$ $(\mu = 1, \ldots, D)$ of the matrices F and G in Eq.(4) are drawn independently across neurons i from a given distribution

[58, 60, 61, 64–68]. In our framework, this distribution is (non-uniquely) parameterised by the density of neuronal locations $\boldsymbol{z}_i$ in the similarity space, and by the choice of the functions $\boldsymbol{f} = (f_1, \ldots, f_D)$ and $\boldsymbol{g} = (g_1, \ldots, g_D)$. For instance, a Hopfield network with P binary patterns [64, 65] corresponds to locations $\boldsymbol{z}_i$ uniformly distributed on the vertices $\{\pm 1\}^P$ of the P-dimensional hypercube and identity functions $f_\mu(\boldsymbol{z}_i) = g_\mu(\boldsymbol{z}_i) = z_{i,\mu}$ [69].

Thus, the assumption of Eq.(3) is a generalised formulation of the connectivity of large low-rank networks. This observation leads to our first result: models of large low-rank networks can generally be represented as circuit models, where each neuron is assigned a location in an underlying similarity space and the connectivity depends on the respective locations of the neurons as expressed by Eq.(3). If the locations $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$ of the neurons are drawn independently, the network dynamics of Eq.(2) then converges to the solution of a neural field equation in the large N limit [61, 70, 71], where the time-dependent membrane potential of each neuron i depends on its location $\boldsymbol{z}_i$. (This convergence result actually holds for *all* networks with $1/N$ scaling of synaptic weights [72].) Therefore, we sometimes refer to a circuit model with many neurons as a "field model" in the rest of the paper.

For the ease of exposition – and without loss of generality (Methods 1) – we consider in the following that the rank D of the connectivity equals the dimensionality K of the input and that the functions $(f_1, \ldots, f_D)$ and $(f_1^{\text{ext}}, \ldots, f_K^{\text{ext}})$ are the same. Our second theoretical result is that, if neuronal locations $\boldsymbol{z}_i$ are sampled independently with probability distribution $\rho(\boldsymbol{z})$, then, in the limit of a large number of neurons $N \to \infty$, the potentials $h_1(t), \ldots, h_N(t)$ can be written as

$$h_i(t) = \sum_{\mu=1}^{D} f_\mu(\boldsymbol{z}_i)\kappa_\mu(t) \tag{5}$$

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

$$\frac{\mathrm{d}}{\mathrm{d}t}\kappa_\mu(t) = -\frac{1}{\tau}\kappa_\mu(t) + \int_{\mathcal{M}} g_\mu(\boldsymbol{z})\, \phi\Big(\sum_{\nu=1}^{D} f_\nu(\boldsymbol{z})\kappa_\nu(t)\Big)\rho(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \; + I_\mu(t). \tag{6}$$

Equation (6) tells us three things. First, it is deterministic, implying that large networks of stochastic LNP neurons follow deterministic collective dynamics [61]. Second, the collective dynamics can be reduced to the nonlinear system of D ordinary differential equations given by Eq.(6), which means that the neural manifold is at most D-dimensional. Third, Eq.(6) expresses the dynamics of the latent variables $\kappa_\mu(t)$, representing the coordinates of the population activity in the D-dimensional neural manifold, in terms of an integral over the d-dimensional similarity space $\mathcal{M}$. This provides an explicit mathematical link between the circuit structure, summarised by the locations of the neurons in the similarity space, and the flow of the collective dynamics on the neural manifold.

It is worth noticing that the low-rank connectivity matrix of Eq.(3) could be the result of Hebbian learning [64, 66]. 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.(3) will build up. In a memory-related brain area, the "experimental condition" could correspond to the presentation of a novel memory item; in early visual cortex, the experimental condition could correspond to elongated spatial edges observed during the critical period [73].

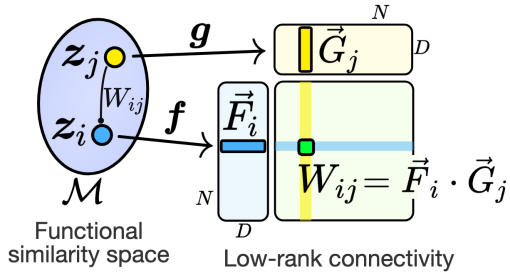

Fig. 2 *Linking circuit structure to low-rank connectivity.* A postsynaptic neuron with location $\boldsymbol{z}_i$ and a presynaptic neuron with location $\boldsymbol{z}_j$ in the functional similarity space $\mathcal{M}$ (left) are mapped by two functions, $\boldsymbol{f}, \boldsymbol{g} : \mathcal{M}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

Circuit structure and neural manifold can mismatch: examples

In this section, 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 transfer function ϕ is a step function of the membrane potential h, i.e. $\phi(h) = R$ for $h > 0$ and zero otherwise. Qualitatively, results do not change for other choices of saturating transfer functions ϕ (Appendix B).

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

In a first toy example, imagine that the functional properties of each neuron i are summarised by a single variable $z_i \in \mathbb{R}$, which is normally distributed across neurons (Fig. 3A1). This determines each neuron's location in the one-dimensional similarity space $\mathcal{M} = \mathbb{R}$. 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) [60, 67]. We study a variant of this model where synaptic weights are given by:

$$W_{ij} = \begin{cases} J\, z_i z_j & (z_i z_j \geq 0) \\ a\, J\, z_i z_j & (z_i z_j < 0) \end{cases} \tag{7}$$

where $J > 0$ is the interaction strength, and negative weights (i.e. inhibitory synapses) are rescaled by a factor $0 < a < 1$. The weights of Eq.(7) can be written in the form of Eq.(3), as: $W_{ij} = f_1(z_i) g_1(z_j) + f_2(z_i) g_2(z_j)$ (Fig. 3A2), meaning that the connectivity has rank $D = 2$. The theory predicts two-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 Methods 2.1). This is confirmed by simulations of a SNN with $N = 2000$ neurons (obeying Eq.(1)) with weights given by Eq.(7). Thus, this model generates a neural manifold of dimension $D = 2$, whereas single-neuron responses depend on their abstract location in a $(d = 1)$-dimensional similarity space.

Two different circuit structures generating identical limit-cycle dynamics

We now turn to an example of limit-cycle dynamics, which is of practical relevance across different brain areas and species. For instance, in a task where a monkey has to grasp a lever and cycle with

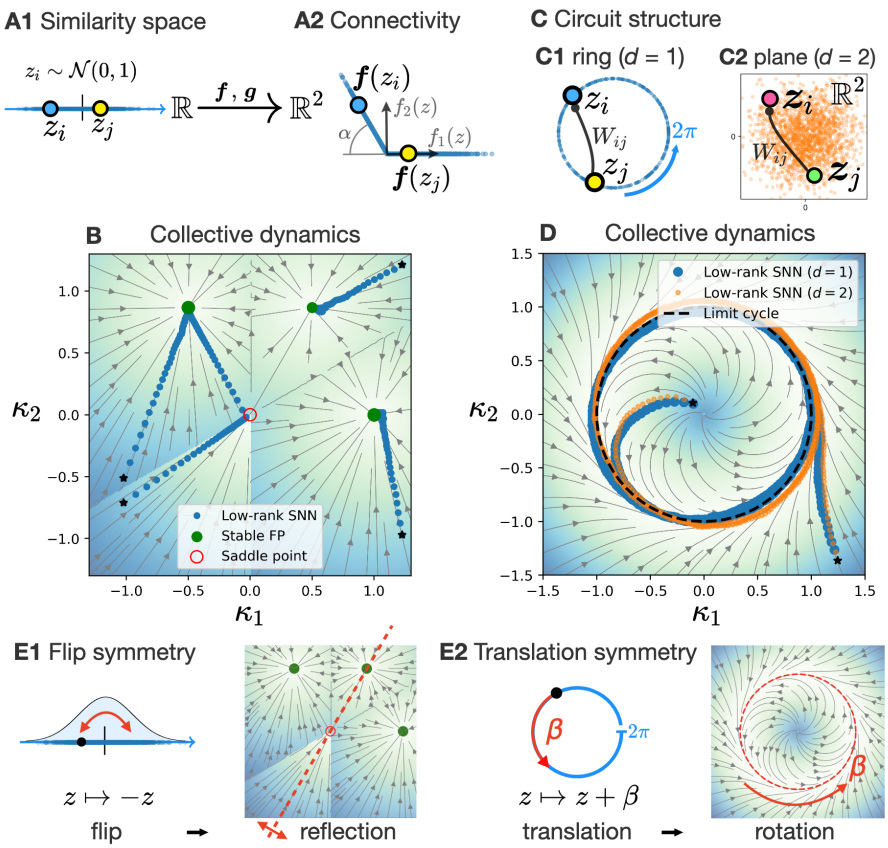

Fig. 3 *The dimension d of the functional similarity space and the dimension D of the neural manifold do not always match.* **A-B.** One-dimensional circuit structure generating a two-dimensional neural manifold. **A1.** Neuronal locations are normally distributed over the real line: the dimension of the similarity space is $d = 1$. **A2.** The connectivity of Eq.(7) is given by 2-dimensional functions $\boldsymbol{f}$ and $\boldsymbol{g} = J\boldsymbol{f}$: the dimension of the neural manifold is $D = 2$. The angle of the kink is $\alpha = \arccos a$. **B.** Flow of the collective dynamics (Eq.(6)) of the circuit model in panel (A) (darker blue-shading indicates faster dynamics). The population trajectories of an SNN with $N = 2000$ neurons (blue dots), starting from 4 different initial conditions (black stars), follow the flow of the collective dynamics. The units of the axes are $k := \frac{\tau JR}{\sqrt{2\pi}}$. **C-D.** Two different circuit models generating identical limit-cyle collective dynamics. **C.** Schematic of the distribution of neuron locations: **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.(8)-(9). The collective dynamics generate a globally attractive limit cycle. The units of the axes are $\varrho^* := \frac{\tau JR}{\pi}\cos\delta$. **E.** *Microscopic* symmetries of the circuit structure induce *macroscopic* symmetries of the collective dynamics. **E1.** The microscopic symmetry w.r.t. a simultaneous flip of all neuron locations (flip symmetry) in panel (A) causes a reflection symmetry of the collective dynamics in panel (B). **E2.** The microscopic symmetry w.r.t. a simultaneous shift 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: $\tau JR = 10$, and $a = 1/2$ in (A-B), $\delta = \pi/10$ in (C-D) (Eqs.(8)-(9)). The low-dimensional trajectory of the SNN is obtained from Eq.(1.8): $\boldsymbol{\kappa}^{\mathrm{SNN}}(t) := \frac{1}{N}\sum_i M\boldsymbol{f}(z_i)h_i(t)$.

its hand, the activity in the motor cortex follows limit-cycle collective dynamics [30] that are likely generated by recurrent mechanisms [44]. In a different context, slow, internally-generated cycling dynamics have been recently observed in the mouse medial enthorinal cortex (MEC) [22] (Fig. 1D).

We build here two models with different circuit structures that generate identical limit-cycle dynamics in a neural manifold of dimension $D = 2$. The first model is a *ring model*, in which neurons are uniformly assigned angular positions $z_i \in [0, 2\pi[$ (Fig. 3C1) [16, 51]. The similarity space thus has dimension $d = 1$. We express the connectivity as:

$$W_{ij} = J \cos(z_i - z_j - \delta) \tag{8}$$

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 [51], whereas for $\delta \neq 0$, the attractor states are travelling bumps [16]. Using the sum expansion of the cosine function [59], we see that the connectivity has rank $D = 2$, where the functions in Eq.(3) 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, neurons have two-dimensional locations $\boldsymbol{z}_i = (z_{i,1}, z_{i,2}) \in \mathbb{R}^2$ normally distributed over the plane (Fig. 3C2); thus, the similarity space has dimension $d = 2$. As in the ring model above, the connections are maximal for an angle δ between two neurons:

$$W_{ij} = J' \, \boldsymbol{z}_i R^{(\delta)} \boldsymbol{z}_j \tag{9}$$

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.(9) has also rank 2, since it is readily expressed as a two-dimensional dot product.

Remarkably, for $J = \sqrt{\frac{\pi}{2}} J'$ (i.e. up to a rescaling), the two models yield *identical* two-dimensional collective dynamics (Methods 2.1). The theory predicts that for any initial condition, the two latent variables that parameterise the population activity converge to a globally attractive limit cycle with angular speed $\tau^{-1} \tan(\delta)$, which is confirmed by simulations of SNNs with $N = 2000$ neurons (Fig. 3D).

We highlight two aspects of this example. (i) 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. More generally, we found that a mismatch between the dimension of the similarity space and that of the neural manifold can originate from heterogeneity across the neuronal population: additional single-neuron properties, parameterising heterogeneity across neurons, can expand the dimensionality of the similarity space without affecting the collective dynamics. For instance, the above "planar" limit-cycle model can be interpreted as a ring model where each neuron is assigned an additional "radial" coordinate, independent of its angular location (Methods 2.2). Nevertheless, we present later, in the section "The footprint of circuit structure", a method that allows us to extract circuit structure from neuronal activity. (ii) For both the "ring" and "planar" models, the connectivity is invariant under the simultaneous rotation of all neurons by any angle. In turn, the collective dynamics exhibits a rotation symmetry, visible in Fig. 3D. The general link between these symmetries is explained in the next section.

Symmetries of circuit structure induce symmetries of collective dynamics

The general notion of a *symmetry* of a dynamical system refers to a transformation of the system that does not affect its dynamics. Our theory links circuit models and neural manifolds on the level

of symmetries: symmetries of a circuit model, expressed as transformations of the neuronal locations in the similarity space, generate symmetries of the collective dynamics in the neural manifold. While this conceptual link is intriguing, its biological relevance is, at this stage, unclear. Nonetheless, in the following, we briefly present our main result and how it applies to the example models above.

Consider a simultaneous transformation of all neuronal locations, expressed as a mapping $S : \mathcal{M} \to \mathcal{M}$ over the similarity space that takes each neuron i from location $\boldsymbol{z}_i$ to location $S\boldsymbol{z}_i$. We call this transformation a *microscopic symmetry* if it commutes with the (microscopic) network dynamics of Eq.(2). Analogously, a *macroscopic symmetry* is a transformation of the latent variables $(\kappa_1, \ldots, \kappa_D)$, expressed as a mapping $T : \mathbb{R}^D \to \mathbb{R}^D$, that commutes with the collective dynamics of Eq.(6). In Methods 3 (and Appendix C), we show that each microscopic symmetry S of a circuit model induces a macroscopic symmetry $T^{(S)}$ of the associated collective dynamics in the D-dimensional neural manifold. For instance, the connectivity of the ring model, Eq.(8), is invariant under the translation of all neurons along the ring by any arbitrary (angular) distance β. This *microscopic* symmetry of the ring model is the origin of the evident *macroscopic* rotational symmetry of its collective dynamics (Fig. 3E2). The earlier example of a one-dimensional circuit model with connectivity given by Eq.(7) also has a microscopic symmetry, given by the "flip" $S : z \mapsto -z$ of each neuron's position. The resulting macroscopic symmetry is the reflection in $\mathbb{R}^2$ w.r.t. the axis spanned by the vector $(1 - a, \sqrt{1 - a^2})$ (Fig. 3E1).

The footprint of circuit structure

We conclude from the previous sections that, despite the link provided by the symmetries, the functional similarity space and the neural manifold can have different dimensionality, and that several circuit models can generate the same collective dynamics.

Yet, even if the circuit structure provably can not be identified from the collective dynamics, can it be recovered by analysing neuronal activity? In this section, we propose an approach to this problem, that we apply on a model of context-dependent decision-making.

Single-neuron statistics reflect neuronal locations in the similarity space

Suppose that we have spiking data from an unknown circuit model network with N neurons. Using standard tools, we can extract at each time point the N-dimensional vector of the population activity, and trace out its trajectory on a D-dimensional neural manifold [27–29, 31–34]. On the other hand, in any circuit model with many neurons (i.e. a field model), the activity of each neuron i depends on its location in the functional similarity space. Thus, two neurons i and j at neighbouring locations $\boldsymbol{z}_i \approx \boldsymbol{z}_j$ will exhibit similar activity.

In order to extract similarity relations between neurons, we choose T different statistical measures that we apply to the spike train of each neuron: this yields an embedding of the neuron in T-dimensional space. Importantly, neighbourhood in the similarity space will lead to similar embeddings. The embeddings associated with several neurons will then reflect their relative positions in the similarity space. Consequently, if sufficiently many neurons are recorded, the d-dimensional similarity space will manifest itself as a point-cloud (with intrinsic dimension d) in the T-dimensional space of statistics. Thus, the above analysis should enable us to extract from N-dimensional spike data a d-dimensional similarity space, in which each neuron is assigned an empirical location.

In principle, neuronal embeddings could be generated using any well-chosen set of T statistical measures that summarise each neuron's activity: for instance, it could be given by each neuron's

tuning to T different population-level factors; or by the "loadings" of the first T principal components (PCs) of population activity associated with each neuron [22] (as we illustrate in the next section). As a key property, these statistics should reflect functional similarities, so that neurons close to (far from) each other in the similarity space generate similar (different) outputs.

As an illustration, we first applied this method to extract the circuit structure of two different SNNs generating similar limit-cycle dynamics, with connectivity given as in the previous examples by Eqs.(8) and (9). We characterised each neuron's spiking activity by its tuning to the two first PCs of the population activity, yielding a two-dimensional embedding of the whole similarity space (Methods 4 and Supp. Fig. S2).

Retrieving circuit structure in models of context-dependent decision-making

Finally, we turned to a more relevant setup for systems neuroscience, in which model networks have to process external inputs to solve a given behavioural task. We applied our method for extracting the similarity space to simulated data obtained from two SNNs which solved the same task. This illustrates: (i) how our framework, linking circuit models and collective dynamics, can be used to build transparent models that solve a given task; and (ii) how different circuit models can be characterised by extracting their similarity space from the neuronal activity.

The chosen task is a simplified instance of context-dependent decision-making [34]. During a given trial, the network receives a sensory input consisting of two independent signals, named 'color' and 'motion', which can both be either positive or negative; as well as a binary context input which indicates which of the 'color' or 'motion' signal is relevant (Fig. 4A). 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. The 'choice' of the network is given by a weighted readout of its low-passed spiking activity.

We designed two SNNs with different circuit structures, composed of $N = 3000$ neurons each, that we will call the "blue" and the "orange" SNN, which both solve the task through similar two-dimensional collective dynamics [58] (Methods 4). For both models, population activity decays towards a baseline state between two trials, while in the 'color' context, symmetric attractive fixed points appear along a 'color'-relevant axis, and similarly in the 'motion' context (Fig. 4B). We aim to identify the circuit structure of each of the two SNNs based on their activity.

Extracting circuit structure from neuronal activity.

The two-dimensional trajectory corresponding to the population activity of each SNN across four trials was obtained by projecting its low-passed, N-dimensional spiking activity onto 'color' and 'motion'-relevant axes [34]. Upon receiving the same sensory and context inputs, the low-dimensional trajectories and the choice readouts of the two SNNs were similar (Fig. 4C,E).

A 3-dimensional embedding of each neuron was then obtained by measuring, for each neuron, the associated loadings onto the first three principal components (PCs) of the population activity (Fig. 4F). For the "blue" SNN, the point-cloud of the embeddings of all the neurons in the population forms a ring; whereas for the "orange" SNN, it lies on a curved, 2-dimensional surface. For both networks, the similarity space inferred from statistical features reflects the underlying circuit structure: the "blue" SNN corresponds to a ring model, and the "orange" SNN to a model with a 2-dimensional similarity space (Methods 4). Neurons were then assigned empirical locations based on their relative positions in the inferred similarity space. A representation of the neuronal activity

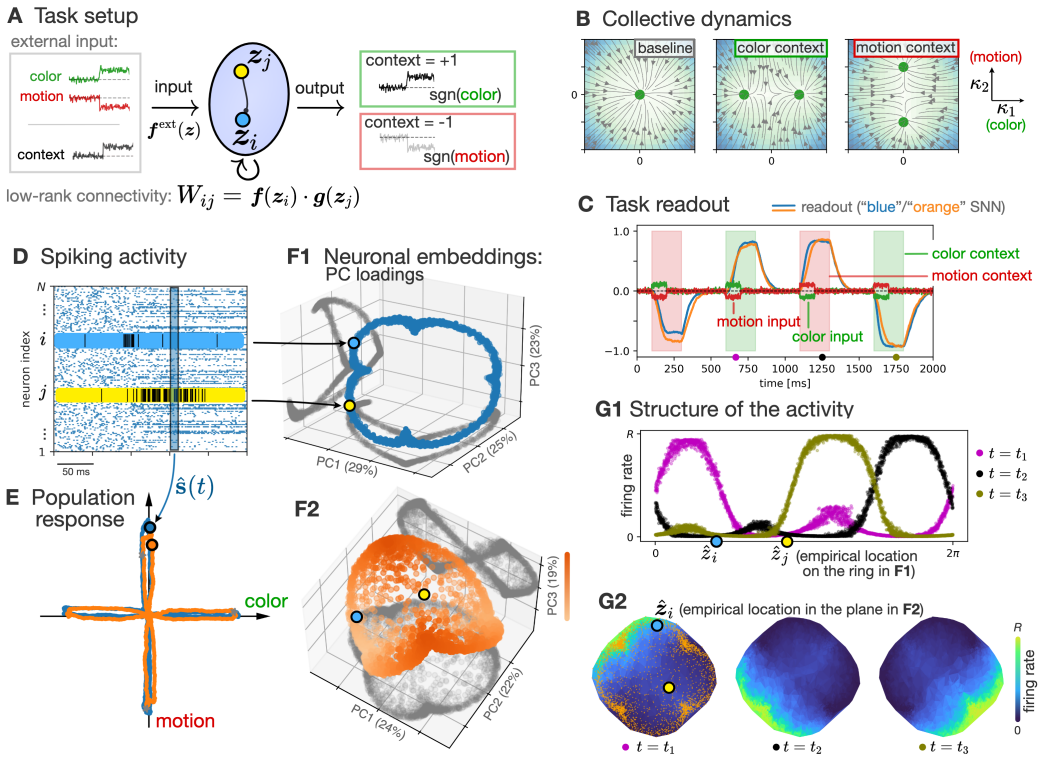

Fig. 4 *Extracting the circuit structure of two SNNs performing the same task.* **A.** During each trial, networks receive an external input consisting of two sensory features (named 'color' and 'motion'), each fluctuating around a positive or negative mean value, and a binary context feature. The task 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 the neurons' locations $\{\boldsymbol{z}_i\}$ in an underlying similarity space. **B.** The task is solved by two-dimensional collective dynamics, where the context input generates attractors (green dots) along an axis corresponding to the relevant sensory input feature. This is implemented in two SNNs, called "blue" and "orange", of $N = 3000$ neurons each **C.** Time-course of the sensory inputs (green and red traces for the 'color' and 'motion' features) and normalised readout of the two SNNs (blue and orange traces) during 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. **D.** Spike trains of 150 example neurons from the "blue" SNN. Spike trains of example neurons i and j are highlighted. **E.** Low-dimensional trajectory of the two SNNs (blue and orange traces) during 4 trials, obtained by projecting the N-dimensional low-passed spiking activity $\hat{\mathbf{s}}(t)$ onto 'color'- and 'motion'-relevant axes. **F.** The spiking activity of each neuron is characterised by the corresponding loadings on the three first principal components (PCs) of the population activity, which determines an embedding of the neuron in three dimensions. **F1.** For the "blue" SNN, the point-cloud of the embeddings of all the neurons in the population forms a ring, revealing the similarity space of the circuit model. **F2.** In contrast, for the "orange" SNN, the set of all three-dimensional embeddings lies on a curved, two-dimensional surface. Points are coloured according to their height (loading on PC3). **G1.** Each neuron i in the "blue" SNN is assigned an empirical location $\hat{z}_i$ according to its position on the ring-like point-cloud in panel F1. Coloured curves represent the neurons' momentary firing rates as a function of their empirical locations, at three different time points $t_{1,2,3}$ (coloured dots on the x-axis of panel C). **G2.** Each neuron i in the "orange" SNN is assigned a two-dimensional empirical location $\hat{\boldsymbol{z}}_i$ according to its position on the two-dimensional point-cloud in panel F2. Heatmaps represent the neurons' firing rates as a function of their empirical locations, at the same three time points $t_{1,2,3}$ (from left to right). Orange dots on the left-most heatmap represent all empirical locations inferred from panel F2.
In panels F1-2, the percentage of variance of the population activity explained by each PC is indicated on the corresponding axis. Grey point-clouds show projections of the points onto the horizontal and a vertical plane. Details about implementation in Methods 4.2-4.3.

as a function of the empirical neuronal locations highlights its underlying spatial structure, which differs between the two networks (Fig. 4G).

Extracting a task-irrelevant circuit structure component

In the previous example, the difference between the circuit model 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 input features 'color' and 'motion'. In both models, individual neurons are tuned to mixtures of these features ("mixed selectivity" [14]). Yet for the "blue" network, the tuning of each neuron to these two features is correlated, so that the corresponding set of two-dimensional tuning vectors lies on a circle; which is not the case for the "orange" network.

Because our approach is task-agnostic, a circuit structure component that is irrelevant to the task at hand can also be extracted. As an illustration, we considered a third "pink" SNN, where the ring circuit structure of the "blue" SNN described above was modified by the addition of a task-irrelevant, Hebbian-like connectivity component (Fig. 5A). To do this, a task-irrelevant feature $\xi_i = \pm 1$ was assigned to each neuron i (chosen independently across neurons with equal probability of being $+1$ or -1); and a term proportional to $\xi_i \xi_j$ was added to the synaptic weight between pairs of neurons i and j (Methods 4). The functional similarity space of this new "pink" model now consists of two rings: each neuron i is jointly characterised by an angular position z_i, and by the binary value of the additional feature ξ_i. Since this additional feature is irrelevant to the task, it cannot be detected through the tuning of the neurons to the task variables (context, color or motion). If both SNNs receive identical external inputs, their respective population activity follow similar low-dimensional trajectories along task-relevant axes (Fig. 5B).

Low-dimensional embeddings of the neurons in both networks were obtained as in Fig. 4 by measuring, for each neuron, the associated loadings onto the first PCs of the population activity. The resulting point-clouds reflect the underlying circuit structure: for the "pink" SNN, the point-cloud forms two disjoint rings, revealing the topology of the similarity space (Fig. 5C); whereas one recovers the original single-ring structure for the "blue" SNN.

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 population activity, we formulated three specific questions in the Introduction. The first question focused on the mathematical link between the structured connectivity of a circuit model and the flow of its collective dynamics in the neural manifold. To answer this question, we proposed a modelling framework that explicitly provides this link. 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 answered the question of whether "low-rank RNNs involve mechanisms similar to classical circuit models or implement truly novel solutions" [3].

The second question, addressing whether a unique relationship exists between the functional similarity space and the neural manifold, must be answered negatively: different circuit structures can yield equivalent collective dynamics. On the other hand, we proved that symmetries of the circuit structure give rise to symmetries of the collective dynamics, a result that highlights the importance of symmetries for models of computation in recurrent networks.

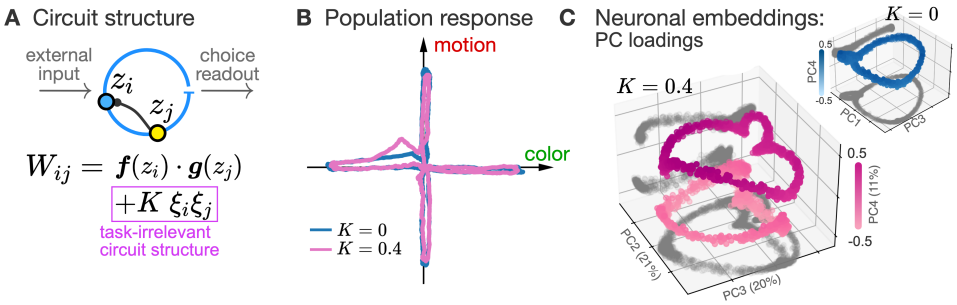

Fig. 5 *Extracting a task-irrelevant circuit structure component from population activity.* **A.** Two SNNs (of $N = 3000$ neurons each) solve the same task as depicted in Fig. 4A. The first SNN corresponds to a ring model (the "blue" SNN of Fig. 4). In the second "pink" SNN, in addition to its location z_i on the ring, each neuron is assigned a task-irrelevant binary feature $\xi_i = \pm 1$, distributed independently across neurons; and a task-irrelevant component $K\xi_i\xi_j$, of strength $K = 0.4 > 0$, is added to the synaptic weights between pairs of neurons i, j. **B.** Low-dimensional trajectories of the two SNNs (blue solid line for the original "blue" SNN, for which $K = 0$; and pink solid line for the "pink" SNN with $K = 0.4$) across 4 trials during which they received the same external input, obtained by projecting the N-dimensional spiking activity on two task-relevant axes (as in Fig. 4D-E). **C.** 3-dimensional embeddings of the neurons in the "pink" ($K = 0.4$) and the "blue" SNN ($K = 0$; small inset). For the "pink" SNN, the three-dimensional point-cloud forms two disjoint rings, revealing the similarity space of the circuit model. Points on the lower or upper ring correspond to neurons with a task-irrelevant feature $\xi_i = -1$ (resp., $\xi_i = +1$). The embedding of each neuron is given by the corresponding loadings on the PCs 2,3 and 4 of the population activity (PCs 1,3,4 for the "blue" SNN). Points are coloured according to their projection onto the vertical axis. Plotting conventions are as in Fig. 4F. Inset: the single-ring circuit structure of the original "blue" SNN is recovered for $K = 0$ (similar to Fig. 4F1).

The third question explored the possibility of extracting knowledge about the circuit structure from recordings of neuronal activity. As an answer to this question, we proposed a method for extracting the topology of the similarity space of a circuit model from the neuronal activity, and applied it on simulated data. Our method allowed us to distinguish between different circuit structures giving rise to the same collective dynamics and solving the same behavioural task. In contrast with the analysis of neuronal tuning or selectivity, this approach is task-agnostic, which allowed us to detect a circuit structure irrelevant to a given behavioural task. In conjunction with task- and behaviour-informed methods [32], the unsupervised approach we propose for characterising functional similarities could enable the discovery of intricate features that modulate neuronal responses in higher cortical areas, and could also be applied in the absence of any task on recordings of freely behaving animals [74, 75].

While our modelling framework provides a link between circuit structure and low-dimensional neuronal activity, it does not incorporate important features of more realistic models of biological neural networks. First, in our model, recurrent connectivity and neuronal properties are constructed according to a homogeneous structural principle (that of field models), whereas cortical neuronal responses exhibit a large degree of trial-to-trial variability [76, 77], which could be explained by chaotic or chaos-like activity due to randomness in the connectivity [58, 66, 78]. The generalisation of our theory to the case of mixed structured and random connectivity, following [58], is left for future work. Second, the neurons in our model are linear-nonlinear-Poisson neurons, which do not account for neuronal refractoriness. Our theory can be partly extended to models of renewal neurons (such as the Spike Response Model or Integrate-and-Fire models [79, 80]). Because of the reset, neuronal dynamics is no longer memory-free and requires an extended formalism [79]. Yet, in the regime of asynchronous firing, each neuron's firing rate can be approximated by a non-linear function of its

synaptic input [81], 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 [82, 83]. This has been taken into account in models of balanced networks [84–87], which can include a spatial structure [88, 89]. Whether our formalism can be extended to derive a low-dimensional description of spatially structured balanced networks remains an open problem.

Finally, we did not apply our method to extract the functional similarity space from experimental data. Yet, in ref. [22], an approach similar to that proposed in this paper was used on recordings of the mouse medial enthorinal cortex: single-neurons were characterised by a few statistics (the PC loadings) extracted from the population activity, and assigned a location on a ring according to these statistics (similar to Supp. Fig. S2); this revealed a bump of neuronal activity slowly cycling over this virtual ring. The question of whether such a low-dimensional population structure can be generally observed in experimental recordings remains to be tested. A positive answer would mean that, even in a population where neurons exhibit mixed selectivity and seemingly heterogeneous activity, neuronal responses can be parameterised by only a few variables, making field models well-suited for describing collective neuronal activity. More broadly, the link between neural manifolds and circuit (or field) models established in this work could potentiate the use of large-scale recordings to uncover circuit structures underlying population-level computation in large networks of neurons.

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] Skaggs, W., Knierim, J., Kudrimoti, H., McNaughton, B.: A model of the neural basis of the rat's sense of direction. In: Tesauro, G., Touretzky, D., Leen, T. (eds.) Advances in Neural Information Processing Systems, vol. 7. MIT Press, ??? (1994)

[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] Gonzalo Cogno, S., Obenhaus, H.A., Lautrup, A., Jacobsen, R.I., Clopath, C., Andersson, S.O., Donato, F., Moser, M.-B., Moser, E.I.: Minute-scale oscillatory sequences in medial entorhinal cortex. Nature **625**(7994), 338–344 (2024)

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

[24] 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)

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

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

[27] 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)

[28] 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)

[29] 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)

[30] Russo, A.A., Bittner, S.R., Perkins, S.M., Seely, J.S., London, B.M., Lara, A.H., Miri, A., Marshall, N.J., Kohn, A., Jessell, T.M., Abbott, L.F., Cunningham, J.P., Churchland, M.M.: Motor cortex embeds muscle-like commands in an untangled population response. Neuron **97**(4), 953–9668 (2018)

[31] Cunningham, J.P., Yu, B.M.: Dimensionality reduction for large-scale neural recordings. Nature Neuroscience **17**(11), 1500–1509 (2014)

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

[33] 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)

[34] 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)

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

[36] 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)

[37] 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)

[38] 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)

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

[40] Gao, P., Ganguli, S.: On simplicity and complexity in the brave new world of large-scale neuroscience. Curr. Op. Neurobiol. **31**(148-155) (2015)

[41] Gao, P., Trautmann, E., Yu, B., Santhanam, G., Ryu, S., Shenoy, K.V., Ganguli, S.: A theory of neural dimensionality, dynamics and measurement. bioRxiv (2017). DOI 10.1101/214262

[42] 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)

[43] 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)

[44] 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)

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

[46] Zugaro, M.B., Arleo, A., Berthoz, A., Wiener, S.I.: Rapid spatial reorientation and head direction cells. Journal of Neuroscience **23**(8), 3478–3482 (2003)

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

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

[49] 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)

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

[51] 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)

[52] 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)

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

[54] 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)

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

[56] 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)

[57] 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)

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

[59] 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)

[60] 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)

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

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

[63] 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)

[64] 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)

[65] 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)

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

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

[68] 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)

[69] 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)

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

[71] 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)

[72] Jabin, P.-E., Schmutz, V., Zhou, D.: Non-exchangeable networks of integrate-and-fire neurons: spatially-extended mean-field limit of the empirical measure. arXiv:2409.06325 (2024)

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

[74] Zong, W., Obenhaus, H.A., Skytøen, E.R., Eneqvist, H., Jong, N.L., Vale, R., Jorge, M.R., Moser, M.-B., Moser, E.I.: Large-scale two-photon calcium imaging in freely moving mice. Cell **185**(7), 1240–125630 (2022)

[75] Manley, J., Lu, S., Barber, K., Demas, J., Kim, H., Meyer, D., Traub, F.M., Vaziri, A.: Simultaneous, cortex-wide dynamics of up to 1 million neurons reveal unbounded scaling of dimensionality with neuron number. Neuron **112**(10), 1694–17095 (2024)

[76] 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)

[77] Stringer, C., Michaelos, M., Tsyboulski, D., Lindo, S.E., Pachitariu, M.: High-precision coding in visual cortex. Cell **184**(10), 2767–277815 (2021)

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

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

[80] 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)

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

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

[83] 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)

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

[85] 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)

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

[87] Ahmadian, Y., Miller, K.D.: What is the dynamical regime of cerebral cortex? Neuron **109**(21), 3373–3391 (2021)

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

[89] Ebsch, C., Rosenbaum, R.: Spatially extended balanced networks without translationally invariant connectivity. The Journal of Mathematical Neuroscience **10**(1), 8 (2020)

[90] 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)

[91] 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)

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

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

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

Methods

1 A unifying framework: field model of low-rank networks 22

2 Similarity space and neural manifold can mismatch 25

3 Symmetries 27

4 Extracting the similarity space 28

1 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 [62]. The spike train $S_i(t)$ of each neuron $i = 1, \ldots, N$ is generated by an inhomogeneous Poisson process with intensity $r_i(t) = \phi(h_i(t))$, 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:

$$\frac{\mathrm{d}}{\mathrm{d}t} h_i(t) = -\frac{1}{\tau} h_i(t) + \frac{1}{N} \sum_{j=1}^{N} W_{ij} S_j(t) \tag{1.1}$$

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:

$$W_{ij} = \sum_{\mu=1}^{D} F_{i,\mu} G_{j,\mu} = \vec{F}_i \cdot \vec{G}_j \tag{1.2}$$

where the dot denotes the dot product in $\mathbb{R}^D$. Each entry W_{ij} of the low-rank connectivity matrix corresponds to the dot 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 (Fig. 2). As discussed in the Results section, Eq.(1.2) 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 [58, 60, 61, 64–68]. In what follows, we introduce a general model for the distribution of the connectivity vectors. The resulting class of RNNs encompasses most known models of mean-field low-rank networks. Moreover, for fixed rank D, as $N \to \infty$, it corresponds to a field model over a similarity 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 *functional properties* of the neuron, sampled from a given distribution.

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

$$F_{i,\mu} = f_\mu(\boldsymbol{z}_i),\ G_{i,\mu} = g_\mu(\boldsymbol{z}_i) \quad (\mu = 1, \dots, D) \tag{1.3}$$

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 field $h_0 : \mathcal{M} \to \mathbb{R}$, according to $h_i(0) = h_0(\boldsymbol{z}_i)$.

Eqs.(1.2),(1.3) define the connectivity 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' locations determine the connectivity. The distribution of the neuronal locations, described by $\rho(\mathrm{d}\boldsymbol{z})$, defines an inner product between functions φ, ψ over $\mathcal{M}$:

$$\langle \varphi, \psi \rangle_\rho := \int_{\mathcal{M}} \rho(\mathrm{d}\boldsymbol{z}) \varphi(\boldsymbol{z}) \psi(\boldsymbol{z}) \tag{1.4}$$

In the following, we assume that the connectivity functions f_μ, g_μ ($\mu = 1, \dots, D$), 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-inner product $\langle \cdot, \cdot \rangle_\rho$.

Corresponding field model.

As $N \to \infty$, the RNN defined by Eqs. (1.1,1.2,1.3) is described by a field model: the membrane potential at time t of a neuron at location $\boldsymbol{z}$ is given by the neural field $v_t(\boldsymbol{z})$, which obeys the *field equation*

$$\partial_t v_t(\boldsymbol{z}) = -\frac{1}{\tau} v_t(\boldsymbol{z}) + \int_{\mathcal{M}} \rho(\mathrm{d}\boldsymbol{z}')\, \boldsymbol{f}(\boldsymbol{z}) \cdot \boldsymbol{g}(\boldsymbol{z}')\, \phi(v_t(\boldsymbol{z}')) \tag{1.5}$$

with initial condition $v_0(\boldsymbol{z}) = h_0(\boldsymbol{z})$. It is the continuum version of Eq.(1.1) [63, 70], where the Poisson noise from the spikes vanishes [61, 71]; 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 [71, 90].

Low-dimensional dynamics.

Analogously to a rank-D RNN [60], the dynamics of the neural field can be written as a D-dimensional system [59], 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 (1.5) can thus be written as an ODE in $\mathcal{H}$: [91]

$$\partial_t v_t = -\frac{1}{\tau} v_t + \sum_{\mu=1}^{D} f_\mu \langle g_\mu, \phi \circ v_t \rangle_\rho \tag{1.6}$$

where $(\circ)$ denotes the composition of functions, and $\langle \cdot, \cdot \rangle_\rho$ the ρ-inner product (Eq.(1.4)). The field v_t undergoes an exponential decay, except in the linear subspace $\mathcal{F} := \text{span}\{f_\mu\} \subset \mathcal{H}$. It can thus be written as a linear combination of the basis functions $\{f_\mu\}$, whose coefficients $\kappa_\mu(t)$ are the "latent variables":

$$v_t(\boldsymbol{z}) = \sum_{\mu=1}^{D} f_\mu(\boldsymbol{z}) \kappa_\mu(t) \ + e^{-t/\tau} h_0^\perp(\boldsymbol{z}) \tag{1.7}$$

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.(1.7) onto each basis function f_μ, the latent variables can be written in vector form as:

$$\boldsymbol{\kappa}(t) = M \langle \boldsymbol{f}, v_t \rangle_\rho \tag{1.8}$$

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. (1.6-1.7-1.8) yield D-dimensional collective dynamics for $\boldsymbol{\kappa}$, which can be written compactly:

$$\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{\kappa}(t) = -\frac{1}{\tau} \boldsymbol{\kappa}(t) + \langle \boldsymbol{g}, \phi \circ \big(\boldsymbol{f} \cdot \boldsymbol{\kappa}(t) \big) \rangle_\rho \tag{1.9}$$

under the assumption that $h_0^\perp = 0$. This equation for the collective dynamics (and its more general expression, Eq.(A.13) in the Appendix) is the central ingredient of our theory. The derivation is given in Appendix A.

Neural manifold and attractor states.

Eq.(1.7) 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 *population activity* $\mathbf{r}(t)$, which is the joint vector of all N instantaneous firing rates $r_1(t), \ldots, r_N(t)$. Besides, $\mathbf{r}(t)$ exponentially decays towards the D-dimensional neural manifold

$$\{\phi(F \cdot \boldsymbol{k}) \mid \boldsymbol{k} \in \mathbb{R}^D\} \subset \mathbb{R}^N \tag{1.10}$$

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.(1.3), 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.(1.9), 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 $I_\nu(t)$ ($\nu = 1, \ldots, K$), and that

the input weight to a neuron i from the input feature I_ν is determined by a corresponding function $f_\nu^{\text{ext}}(\boldsymbol{z}_i)$ of the neuron's location $\boldsymbol{z}_i$. The external input it receives can thus be written:

$$I_i^{\text{ext}}(t) = \sum_{\nu=1}^{K} f_\nu^{\text{ext}}(\boldsymbol{z}_i) I_\nu(t) \tag{1.11}$$

This simply results in a drive to the collective dynamics of Eq.(1.9), while expanding their dimensionality – at most up to $D + K$. The corresponding general expression of the collective dynamics is given by Eq.(A.13) 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 [59]. Both generalisations are given in Appendix A.

2 Similarity space and neural manifold can mismatch

2.1 Examples

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

In our first example, neuronal locations 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 $i = 1, \ldots, N$ (Fig. 3A1). The interactions are given by Eq.(7):

$$W_{ij} = \begin{cases} J\, z_i z_j & (z_i z_j \geq 0) \\ \cos(\alpha) J\, z_i z_j & (z_i z_j < 0) \end{cases} \tag{2.1}$$

with $0 < \alpha < \pi/2$ (and $a = \cos\alpha$ in Eq.(7)). This can be written as the 2-dimensional dot product $W_{ij} = \boldsymbol{f}(z_i) \cdot \boldsymbol{g}(z_j)$, with:

$$\boldsymbol{f}(z) = \begin{pmatrix} [z]_+ + \cos(\alpha)[z]_- \\ -\sin(\alpha)[z]_- \end{pmatrix} \in \mathbb{R}^2 \quad ; \quad \boldsymbol{g} = J\boldsymbol{f} \tag{2.2}$$

where $[z]_\pm = z \cdot \mathbb{1}_{\{\text{sgn}(z)=\pm 1\}}$ denotes the positive or negative part of z. This is illustrated on Fig. 3A2. The corresponding collective dynamics resulting from Eq.(1.9) and displayed in Fig. 3B are derived in Appendix B.1.

Two different circuit structures generate identical limit-cycle dynamics

We consider two circuit models, with connectivity given by Eqs.(8)-(9), which we recall here for convenience. The first model is a ring model, with angular locations z_i uniformly distributed on the ring $[0, 2\pi[$, and rank-2 connectivity given by:

$$W_{ij} = w^{\text{ring}}(z_i, z_j) := J\cos(z_i - z_j - \delta) = J \begin{pmatrix} \cos z_i \\ \sin z_i \end{pmatrix} \cdot \begin{pmatrix} \cos(z_j + \delta) \\ \sin(z_j + \delta) \end{pmatrix} \tag{2.3}$$

with δ an arbitrary phase shift. In the second "planar" model, neurons have two-dimensional locations $\boldsymbol{z}_i = (z_{i,1}, z_{i,2}) \in \mathbb{R}^2$ normally distributed over the plane; and synaptic weights are given by:

$$W_{ij} = J' \, \boldsymbol{z}_i R^{(\delta)} \boldsymbol{z}_j \tag{2.4}$$

where $R^{(\delta)}$ is the 2×2 matrix of rotation by an angle δ in the plane; and $J' > 0$.

Remarkably, for a step transfer function $\phi(x) = R \cdot \mathbb{1}_{\{x>0\}}$, and $J = \sqrt{\frac{\pi}{2}} J'$, the two models obey *identical* collective dynamics. For the polar coordinates $(\varrho(t), \theta(t))$ of the latent variable $\boldsymbol{\kappa}(t) \in \mathbb{R}^2$, Eq.(1.9) yields

$$\dot{\varrho} = -\frac{1}{\tau}\varrho + \frac{J}{\pi} R \cos\delta \quad ; \quad \varrho\dot{\theta} = \frac{J}{\pi} R \sin\delta \tag{2.5}$$

for *both* models (derivation in Appendix B.2).

For any initial condition, the latent variables converge to a globally attractive limit cycle with constant angular speed which only depends on τ and δ (given by $\dot{\theta} = \frac{1}{\tau}\tan\delta$). This last point generalises to arbitrary transfer function ϕ and any rotationally invariant distribution of the neuronal locations (Appendix B.2).

2.2 General interpretation

Heterogeneity as a source of dimensionality mismatch.

We observed that the dimensionality d of the similarity space can be expanded by additional single-neuron properties, that parameterise heterogeneity across the population. As a concrete example, consider the "planar" limit-cycle model above with a two-dimensional similarity space (Eq.(2.4)). In this example, neuronal locations are normally distributed in $\mathbb{R}^2$. Thus, the location $\boldsymbol{z}_i$ of any neuron i can be written in polar coordinates, with an angle φ_i (distributed uniformly over the ring) and a radius ξ_i (distributed following the χ_2-distribution, independent from φ_i). The connectivity of Eq.(2.4) then 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.(2.3)). We now interpret the radius ξ_i as a single-neuron property that parameterises an heterogeneity across the population, affecting the connectivity of the original ring model (for which $\xi_i = 1$ for all neurons). Yet, this heterogeneity does not affect the collective dynamics of the model (Appendix B.2). Building on this observation, we found that 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 all outgoing weights of each neuron by a random factor, that describes variations of synapse strength; and heterogeneous maximum firing rates across neurons (Appendix D).

Several universal approximators.

The mismatch between the circuit structure and the collective dynamics can also be understood from another perspective: several circuit models with different similarity spaces can approximate the same collective dynamics. We identified 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 [60]. This class of models 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.(3) can be chosen arbitrarily with no requirement on their "smoothness" – analogous to the universal approximation property of feed-forward neural networks (Appendix E.2).

3 Symmetries

A *microscopic symmetry* of a circuit model can be informally defined as follows. Consider a transformation of the neuronal locations, expressed as a bijection $S : \mathcal{M} \to \mathcal{M}$ over the similarity space that takes each neuron i from location $\boldsymbol{z}_i$ to location $S\boldsymbol{z}_i$. There are two ways in which this mapping can be combined with the microscopic network dynamics of Eq.(2). Given the set of all membrane potentials at initial time, one can first let the network evolve up to time t and then "move" each neuron i from location $\boldsymbol{z}_i$ to $S\boldsymbol{z}_i$. Alternatively, one can first "move" the neurons and then let the network evolve up to time t. A microscopic symmetry of the model is a transformation S of the neuronal locations for which these two orders of composition are equivalent (Fig. S1).

Analogously, a *macroscopic symmetry* is a transformation of the latent variables $(\kappa_1, \ldots, \kappa_D)$ that commutes with the collective dynamics of Eq.(6).

A symmetry is a mapping that commutes with the flow

A more precise formulation relies on 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 location; i.e., $h_i(0) = v_0(\boldsymbol{z}_i)$. The *flow* of the dynamics transforms this initial condition 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 $v_t = \Phi_t[v_0]$ (Fig. S1A, left panels). On the other hand, if one "moves" each neuron from location $\boldsymbol{z}_i$ to location $S\boldsymbol{z}_i$, its membrane potential is now given by the value of the field at its "new" position; i.e. one now has $h_i(0) = v_0(S\boldsymbol{z}_i)$. In other words, "moving" the neurons is equivalent to a composition of the field with S (Fig. S1A, top panels). As described above, the bijection S is a *microscopic symmetry* of the model if "moving" the neurons before or after the time evolution is equivalent; i.e., $\Phi_t[v_0 \circ S] = \Phi_t[v_0] \circ S$.

Likewise, a *macroscopic* symmetry of the model is a transformation $T : \mathbb{R}^D \to \mathbb{R}^D$ of the latent variables $(\kappa_1, \ldots, \kappa_D)$ that commutes with the flow of the collective dynamics of Eq.(6) (Fig. S1B).

From microscopic to macroscopic symmetries

We show that any microscopic symmetry S of a field model generates a macroscopic symmetry $T^{(S)}$ of its collective dynamics. For the latter, we give an explicit expression under a reasonable additional assumption. More precisely (Theorem 1 in Appendix C):

1. Let S be (i) a microscopic 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 $\nu = 1, \ldots, D$. Then, a symmetry of the collective dynamics is given by the linear mapping $T^{(S)}$, defined by the $D \times D$ matrix

$$T^{(S)}_{\mu\nu} = \sum_\sigma M_{\mu\sigma} \langle f_\sigma, f_\nu \circ S \rangle_\rho \tag{3.1}$$

   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 locations 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.

A fortiori, any symmetry $T^{(S)}$ of the collective dynamics will be a symmetry of the set of attractor states of the population activity.

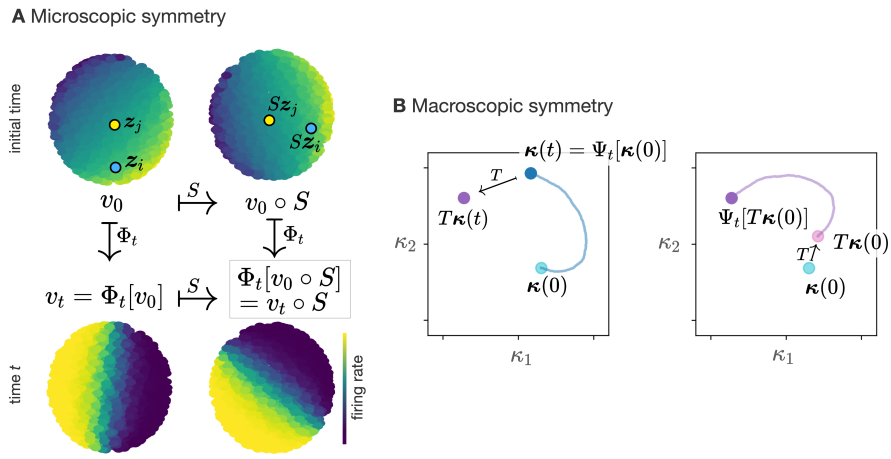

Fig. S1 *A symmetry is a mapping that commutes with the flow.* **A.** Heatmaps represent the field of instantaneous firing rates, $\phi(v_t(z))$, over a disc-like, $d = 2$-dimensional similarity space, at initial time (upper panels) and final time t (lower panels). A *microscopic symmetry* of a field model is a transformation $S : \mathcal{M} \to \mathcal{M}$ of the neuronal locations that commutes with the flow Φ_t of the field dynamics up to any time t, i.e. such that the upper and lower path are equivalent. The upper path (right-down) first composes the initial field v_0 with S, then applies Φ_t, yielding the final state $\Phi_t[v_0 \circ S]$. The lower path (down-right) first applies Φ_t, then composes the field with S, yielding the final state $v_t \circ S = \Phi_t[v_0] \circ S$. **B.** A *macroscopic symmetry* of the model is a mapping $T : \mathbb{R}^D \to \mathbb{R}^D$ of the latent variables that commutes with the flow Ψ_t of the collective dynamics up to any time t: the final state (purple dot) is the same in the left panel (first apply Ψ_t, and then T) and in the right panel (first apply T, and then Ψ_t). Solid lines represent the trajectory of the latent variables in a $D = 2$-dimensional neural manifold, from initial time up to time t, starting from the initial condition $\boldsymbol{\kappa}(0)$ (left panel) or $T\boldsymbol{\kappa}(0)$ (right panel).

Symmetry groups

The mapping $\mathcal{T} : S \mapsto T^{(S)}$ defines the action on $\mathbb{R}^D$ of any group G of *microscopic* symmetries of the model, which will also be a group of *macroscopic* symmetries. 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 $\mathbb{R}^D$ (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 $\mathbb{R}^3$.

4 Extracting the similarity space

4.1 Models of limit-cycle dynamics

As a first illustration of our method for extracting the similarity space from the neuronal activity, we studied two spiking models that generate similar limit-cycle dynamics. SNNs were implemented as in the examples of Fig. 3C-D, with connectivity given by Eqs.(8) and (9); the only difference is that we now consider a sigmoidal transfer function ϕ[1]. Since both models yield similar collective dynamics (Appendix B.2), the activity of the two networks generates similar low-dimensional trajectories, that both enter into a limit cycle in a $D = 2$-dimensional neural manifold (Fig. S2A,B).

[1] so that the field of instantaneous firing rates, given by $\phi(v_t(z))$, is continuous over the similarity space.

To distinguish the two networks from each other, we characterised the spiking activity of each neuron with two statistical measures, chosen to be the neuron's tuning to each of the two first PCs of the low-passed spiking activity. This yields an two-dimensional embedding of each neuron. The point-cloud of the set of all neuronal embeddings enables us to visualise the functional similarity space (up to an isomorphism) and to distinguish between the two different models (Fig. S2C): for the first SNN, the similarity space is a one-dimensional ring (Fig. S2C1); for the second SNN, it is a two-dimensional surface (Fig. S2C2). For both models, the empirical location of the neurons in the similarity space was then obtained by parameterising their relative positions in this point-cloud. A representation of the neuronal activity as a function of the empirical neuronal locations highlights the underlying spatial structure, which differs between the two models (Fig. S2D). Finally, sorting the neurons with respect to their empirical locations makes oscillatory dynamics clearly visible (Fig. S2E), similar to a recent analysis that revealed slow oscillatory dynamics in the mouse MEC [22]. We stress that sorting the neurons is equivalent to arranging them in one-dimensional space; higher-dimensional neuronal embeddings can thus be seen as a generalisation of neuron sorting (Fig. S2,D2).

4.2 Models of context-dependent decision-making

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 location $\boldsymbol{z}_i$, as expressed by Eq.(2). Similarly, the output is a one-dimensional readout of the spiking activity, with weights determined by the neurons' locations.

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 $f_1^\mathrm{ext}(\boldsymbol{z}_i), f_2^\mathrm{ext}(\boldsymbol{z}_i)$ and $f_\mathrm{c}(\boldsymbol{z}_i)$, 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:

$$I(t, \boldsymbol{z}_i) = \boldsymbol{f}(\boldsymbol{z}_i) \cdot \boldsymbol{I}(t) + \frac{1}{\tau} J_\mathrm{c} f_\mathrm{c}(\boldsymbol{z}_i) I_\mathrm{c}(t) \tag{4.1}$$

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

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 (or tuning) 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 their general expression, Eq.(A.13)):

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\kappa}(t) = -\frac{1}{\tau}\boldsymbol{\kappa}(t) + J\langle \boldsymbol{f}, \phi\left(\boldsymbol{f} \cdot \boldsymbol{\kappa}(t) + J_\mathrm{c} f_\mathrm{c} I_\mathrm{c}\right)\rangle_\rho + \boldsymbol{I}(t). \tag{4.2}$$

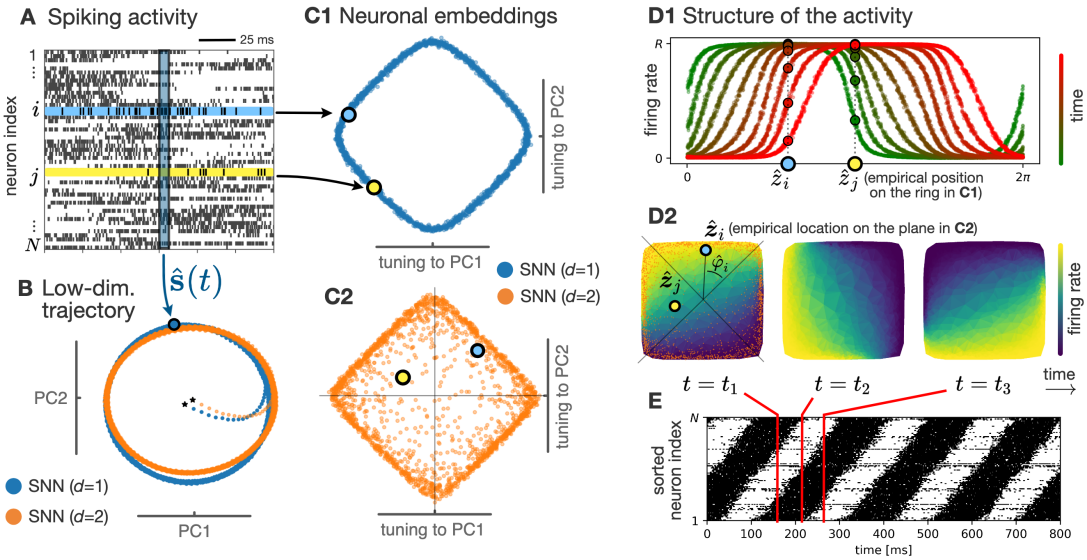

Fig. S2 *The functional similarity space can be extracted from the neuronal activity.* Two SNNs of $N = 2000$ neurons follow similar limit-cycle collective dynamics, even though they have a different cicuit structure (Fig. 3C,D). **A.** Spike trains of randomly chosen example neurons in the SNN with $d = 1$, with two neurons i and j highlighted. At any time t, the N-dimensional population activity lies on the neural manifold. In contrast, the spike train of each neuron i depends on its location $\boldsymbol{z}_i$ in the similarity space. **B.** The trajectory of the two SNNs is shown in the space of the first two principal components (PCs) of the low-passed spiking activity $\hat{\mathbf{s}(\mathbf{t})}$ ($> 80\%$ variance explained for both networks), after a rotation to superimpose the two trajectories. Both trajectories enter a limit-cycle with the same period $T = 195$ ms. **C.** The spiking activity of each neuron was characterised by its tuning to each of the two first principal components (PCs) of the population activity. **C1.** For the first SNN, the point-cloud formed by the embeddings of all the neurons in the population forms a ring. **C2.** In constrast, for the second SNN, the set of all two-dimensional embeddings lies on a two-dimensional surface. For both networks, it is consistent with the similarity space of the underlying circuit model (Fig. 3C). **D1.** Each neuron i of the first SNN is assigned an empirical location $\hat{z}_i$ according to its position on the ring-like point-cloud of panel C1. Coloured curves represent the neurons' firing rates as a function of their empirical locations, at several time points (where colour from green to red represents increasing time). The time series of the firing rate of two neurons i and j are highlighted. **D2.** Each neuron i of the second SNN is assigned a two-dimensional empirical location $\hat{\boldsymbol{z}}_i$ according to its position in the point-cloud of panel C2. Heatmaps represent the neurons' momentary firing rates as a function of their empirical locations, at three time points t_1, t_2, t_3 (from left to right). Orange dots on the left-most heatmap represent all empirical locations inferred from panel C2. The whole point-cloud is displayed with a 45° rotation compared to panel C2 for convenience. **E.** Rasterplot of the spiking activity of the second SNN, where neurons were sorted according to the angle $\hat{\varphi}_i$ of their empirical location (panel D2, left). This makes oscillatory dynamics clearly visible, similar to the analysis performed in [22]. Red bars correspond to the three time points shown in panel D2. Details about implementation in Methods 4.3.

The connectivity functions $\boldsymbol{f}(z)$, the context selectivity $f_c(z)$ and the strength parameters J, J_c were chosen such that in the absence of external input ($I_1 = I_2 = I_c = 0$), these dynamics admit no non-trivial fixed point; but in the 'color' context ($I_c = +1$), symmetric attractive fixed points appear along the 'color' axis at $\boldsymbol{\kappa} = (\pm k, 0)$, with $k > 0$ given by the fixed points of Eq.(4.2) with $I_c = +1$. Similarly, the 'motion' context ($I_c = -1$) yields attractive fixed points along the 'motion' axis at $\boldsymbol{\kappa} = (0, \pm k)$, as illustrated in Fig. 4B. More details about the implementation are given below.

This was 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 $f_{\text{out}}(\boldsymbol{z}_i)$:

$$\text{readout}(t) = \frac{1}{N} \sum_{i=1}^{N} f_{\text{out}}(\boldsymbol{z}_i) \hat{s}_i(t) \tag{4.3}$$

with $\hat{s}_i(t)$ the low-passed 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 relevant sensory input: $(+k, 0)$ in the 'color' context if $I_1 > 0$, or $(0, +k)$ in the 'motion' context if $I_2 > 0$. Similarly, the readout is negative for a negative relevant sensory input.

Low-dimensional trajectory of the SNN

Low-dimensional trajectories of the SNNs in Figs.4E and 5B were obtained by projecting the low-passed N-dimensional spiking activity onto 'color' and 'motion'-relevant 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.(4.3) is equivalent to the projection of this 2-dimensional population response on the diagonal axis $(1, 1)$ (Fig. 4C,E).

Implementation with two circuit models

The first "blue" model in the main text is a ring model: neurons have angular positions uniformly distributed over $[0, 2\pi[$; and the connectivity is given by $W_{ij} = J \cos(z_i - z_j)$ (as in Eq.(8), with $\delta = 0$) which is written in rank-2 form using $\boldsymbol{f}(z) = (\cos z, \sin z)$.

The second "orange" model has a 2-dimensional functional similarity space: neuronal locations are uniformly distributed in the square $[-1, 1]^2$ (i.e., $\boldsymbol{z}_i = (z_{i,1}, z_{i,2})$ where $z_{i,1}$ and $z_{i,2}$ are independent and uniformly distributed in $[-1, 1]$).The connectivity is the dot product of the locations: we set $\boldsymbol{f}(\boldsymbol{z}_i) = \boldsymbol{z}_i$, so that $W_{ij} = J\, \boldsymbol{z}_i \cdot \boldsymbol{z}_j$. For both models, the voltage-to-rate function was a shifted sigmoid: $\phi(x) = R(1 + \tanh(x - 1))/2$. In the absence of sensory external input, Eq. (4.2) admits fixed points given by

$$\boldsymbol{\kappa} = \begin{cases} (\pm k, 0) \text{ if } I_c \geq 0 \\ (0, \pm k) \text{ if } I_c \leq 0 \end{cases} \tag{4.4}$$

where $k \geq 0$ is the solution to:

$$\begin{aligned} \text{“blue” ring model:} \quad & k = \tau J \int_0^{2\pi} \frac{dz}{2\pi} \cos z \; \phi\Big(k \cos z + J_\mathrm{c} |I_\mathrm{c}| (|\cos z| - |\sin z|)\Big) \\ \text{“orange” planar model:} \quad & k = \tau J \int_{-1}^{1} \frac{dz_1}{2} \int_{-1}^{1} \frac{dz_2}{2} \, z_1 \, \phi\Big(k z_1 + J_\mathrm{c} |I_\mathrm{c}| (|z_1| - |z_2|)\Big) \end{aligned} \tag{4.5}$$

The values of the parameters J and J_c were chosen so that the above equations have no solutions $k \neq 0$ for $I_\mathrm{c} = 0$ (corresponding to the “baseline” dynamics in Fig. 4B), but have a non-trivial solution $k > 0$ for $|I_\mathrm{c}| = 1$. Solutions were found numerically. The chosen values were $J = 0.7$, $J_\mathrm{c} = 1.5$ for both models (given the value of $\tau R = 10$).

Additional task-irrelevant connectivity component

We finally consider a modification of the “blue” ring model, where each neuron is assigned an additional feature $\xi_i = \pm 1$, chosen randomly for each neuron i, and the recurrent connectivity is given by:

$$W_{ij} = J \cos(z_i - z_j) + K \xi_i \xi_j \tag{4.6}$$

with $K \geq 0$. Each neuron i is thus characterised by its location on a double-ring, $(z_i, \xi_i) \in [0, 2\pi[\times \{\pm 1\}$. The recurrent connectivity has rank 3, with connectivity functions given by $\boldsymbol{f}(z, \xi) = (\cos z, \sin z, \xi) \in \mathbb{R}^3$, and $\boldsymbol{g}(z, \xi) = (J \cos z, J \sin z, K\xi)$. The collective dynamics are three-dimensional, and read in the absence of sensory external input (similar to Eq.(4.2)):

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{\kappa}_{1:2} &= -\frac{1}{\tau} \boldsymbol{\kappa}_{1:2} + J \langle \boldsymbol{f}_{1:2}, \phi \circ v_t \rangle_\rho \\ \frac{\mathrm{d}}{\mathrm{d}t} \kappa_3 &= -\frac{1}{\tau} \kappa_3 + K \langle \xi, \phi \circ v_t \rangle_\rho \end{aligned} \tag{4.7}$$

where $v_t(z, \xi) = \boldsymbol{f}_{1:2}(z) \cdot \boldsymbol{\kappa}_{1:2}(t) + \xi \kappa_3(t) + J_\mathrm{c} I_\mathrm{c} f_\mathrm{c}(z)$, and $\boldsymbol{\kappa}_{1:2} = (\kappa_1, \kappa_2)$, $\boldsymbol{f}_{1:2} = (f_1, f_2)$.

Notice that if $\kappa_3 = 0$, then v_t does not depend on ξ. Thus $\langle \xi, \phi \circ v_t \rangle_\rho = 0$; and the first line of Eq.(4.7) becomes identical to Eq.(4.2). Therefore, the fixed points of Eq.(4.2) for $\boldsymbol{\kappa}_{1:2}$ (given by Eq.(4.4)) are also fixed points of Eq.(4.7) given $\kappa_3 = 0$. Yet, those fixed points should be stable (in the direction of κ_3), in order for the model to behave similarly to the original “blue” model. The condition of stability in the direction of κ_3 reads (from Eq.(4.7)):

$$\frac{\partial}{\partial \kappa_3} \left[-\frac{1}{\tau} \kappa_3 + K \langle \xi, \phi \circ v_t \rangle_\rho \right]_{\kappa_3 = 0} < 0 \tag{4.8}$$

which yields, after explicit computation:

$$K < \left(\tau \langle \xi^2 \rangle \langle \phi' \rangle \right)^{-1} =: K^* \tag{4.9}$$

where $\langle \xi^2 \rangle$ ($= 1$ in our case) is the average value of ξ^2; and $\langle \phi' \rangle = \int_0^{2\pi} \frac{dz}{2\pi} \phi' \left(k \cos z + J_\mathrm{c} |I_\mathrm{c}| f_\mathrm{c}(z) \right)$ is the average value of the “gain” function ϕ'.

In Fig. 5, K was set near below the critical value K^* at which the fixed points with $\kappa_3 = 0$ lose their stability. In addition, the network received a weak external input in the direction of κ_3, of the

form

$$I^{\text{noise}}(t; z_i, \xi_i) = \zeta(t)\xi_i, \tag{4.10}$$

where $\zeta(t)$ is an Ornstein-Uhlenbeck process with mean 0 and small variance. In the original "blue" model with $K = 0$, the effect on the dynamics was negligible. In the model with K close to the critical value K^*, the additional circuit structure dramatically enhances this weak input, leading to large fluctuations of κ_3, and thus to a strong dependence of the membrane potentials (and thus of the firing rates) on the task-irrelevant features ξ_i.

The value of the critical strength (Eq.(4.9)) was computed to be $K^* \approx 0.67$ for $I_c = \pm 1$, and $K^* \approx 0.48$ for $I_c = 0$ (given the parameters $\tau R = 10$, and $J = 0.7, J_c = 1.5$; which yields from Eq.(4.5) $k \approx 1.29$ for $I_c = \pm 1$, and $k = 0$ for $I_c = 0$). K was then set to 0.4.

4.3 Numerical simulations

In all simulations (Figs. 3, 4, 5, S2), the neuronal membrane time constant was set to $\tau = 10$ ms, which defines all characteristic time scales. The maximum firing rate was $R = 10/\tau$. In Figs. 4, 5, S2, spikes were low-passed with a time constant $\alpha = \tau$, and the spike train of each neuron was then z-scored by substracting its mean and dividing by its standard deviation. PCA was performed on the z-scored low-passed activity.

For the decision-making task, analyses were performed on 8 trials that covered all possible combinations of the 3 binary task variables (mean $\pm\bar{I}$ of the 'color' and 'motion' inputs, and value ± 1 of the 'context' input) (only 4 trials are shown on Figs. 4 and 5). The magnitude of the sensory inputs (see below Eq.(4.1)) was set to $\bar{I} = 0.1\tau^{-1}$, and their standard deviation to $\sigma = \bar{I}/5$.

In Figs. 4 and 5, all three models described above ("blue", "orange" and "pink") received the same low-dimensional input, including the sensory and context features, and also the noisy input with binary random projections ξ_i given by Eq.(4.10). The effect on the dynamics was negligible for models whose connectivity was uncorrelated with ξ_i ("blue" and "orange" SNNs). Parameters of the O.-U. process $\zeta(t)$ were $\theta = 0.1\tau^{-1}$ and $\sigma = \theta/5$.

In Fig. S2, the tuning of each neuron i to the two first PCs was computed as the normalised difference between spike counts of the neuron during periods where each PC was positive or negative:

$$[\text{tuning to PC}k](i) = \frac{1}{T}\int_0^T S_i(t)\left(2 \cdot \mathbb{1}_{\{\text{PC}k(t)>0\}} - 1\right) dt \tag{4.11}$$

where PC$k(t)$ ($k = 1, 2$) represents the trajectory of the k-th principal component of the activity, and $S_i(t)$ the spike train of neuron i. This choice of statistics allowed one to keep information about the variance of each neuron's activity, which was varying across neurons for the SNN with $d = 2$. In contrast, this information was collapsed in the PC loadings associated with the z-scored population activity (due to the z-scoring).

Appendix

A Collective dynamics **34**

B Examples of collective dynamics **37**

C Symmetries **40**

D Adding heterogeneity in a field model **44**

E Field models as universal approximators **45**

Note: in the Appendix, we drop the boldface notation $\boldsymbol{z}$ to denote neuronal locations. 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.(1.6), rewritten here for convenience:

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

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 $\mathcal{F}^\perp$:

$$v_t = \sum_\mu f_\mu \kappa_\mu(t) + v_t^\perp \tag{A.2}$$

which implies:

$$\partial_t v_t = \sum_\mu f_\mu \dot{\kappa}_\mu(t) + \partial_t v_t^\perp \tag{A.3}$$

where the dot ($\dot{}$) denotes the time derivative. Together, equations (A.1)-(A.2) yield:

$$\partial_t v_t = \sum_\mu f_\mu \left(-\frac{1}{\tau} \kappa_\mu(t) + \langle g_\mu, \phi \circ v_t \rangle_\rho \right) - \frac{1}{\tau} v_t^\perp \tag{A.4}$$

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.(A.3) and (A.4):

$$\dot{\kappa}_\mu(t) = -\frac{1}{\tau} \kappa_\mu(t) + \langle g_\mu, \phi \circ v_t \rangle_\rho \tag{A.5}$$

$$\partial_t v_t^\perp = -\frac{1}{\tau} v_t^\perp \tag{A.6}$$

Thus, $v_t^\perp = e^{-t/\tau} h_0^\perp$, so that Eq.(A.2) yields Eq.(1.7). It allows replacing v_t in Eq.(A.5) to obtain, in vector form:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\kappa}(t) = -\frac{1}{\tau}\boldsymbol{\kappa}(t) + \langle \boldsymbol{g}, \phi \circ \Big(\boldsymbol{f}\cdot\boldsymbol{\kappa}(t) + e^{-t/\tau} h_0^\perp\Big)\rangle_\rho \tag{A.7}$$

which yields Eq.(1.9) if we assume $h_0 \in \mathcal{F}$.

Eventually, the expression of the latent variables $\boldsymbol{\kappa}$ can be obtained by projecting Eq.(A.2) 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.(1.8):

$$\boldsymbol{\kappa}(t) = C^{-1}\langle \boldsymbol{f}, v_t\rangle_\rho \tag{A.8}$$

A.2 Including external inputs

Assume that the network receives an external input composed of K features $I_\nu(t)$ ($\nu = 1, \dots, K$), to each of which the tuning of any neuron i is determined by a corresponding function $f_\nu^{\mathrm{ext}}(z_i)$ of its location z_i. The external input received by any neuron can be written: $I^{\mathrm{ext}}(t, z_i) = \sum_{\nu=1}^K f_\nu^{\mathrm{ext}}(z_i) I_\nu(t)$. The total (recurrent + external) input thus now belongs to the linear subspace $\tilde{\mathcal{F}} \subset \mathcal{H}$ spanned by $\{f_\mu\}_{\mu=1}^D \cup \{f_\nu^{\mathrm{ext}}\}_{\nu=1}^K$, 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

$$w(z, z') = \sum_{\mu=1}^{D} f_\mu(z) g_\mu(z') = \sum_{\alpha=1}^{\tilde{D}} \tilde{f}_\alpha(z) \tilde{g}_\alpha(z') \tag{A.9}$$

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

$$I^{\mathrm{ext}}(t, z) = \sum_{\nu=1}^{K} f_\nu^{\mathrm{ext}}(z) I_\nu(t) = \sum_{\alpha=1}^{\tilde{D}} \tilde{f}_\alpha(z) \tilde{I}_\alpha(t) \tag{A.10}$$

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

$$I(t, z) = \sum_{\alpha=1}^{\tilde{D}} \tilde{f}_\alpha(z) \Big(\int_{\mathcal{M}} \tilde{g}_\alpha(z') \phi(v_t(z')) \rho(\mathrm{d}z) \; + \tilde{I}_\alpha(t) \Big) \tag{A.11}$$

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:

$$\partial_t v_t = -\frac{1}{\tau} v_t + \sum_\alpha f_\alpha \Big(\langle g_\alpha, \phi \circ v_t\rangle_\rho + I_\alpha(t) \Big) \tag{A.12}$$

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

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\kappa}(t) = -\frac{1}{\tau}\boldsymbol{\kappa}(t) + \langle \boldsymbol{g}, \phi \circ \left(\boldsymbol{f}\cdot\boldsymbol{\kappa}(t) + e^{-t/\tau}h_0^{\perp}\right)\rangle_\rho + \boldsymbol{I}(t) \tag{A.13}$$

where $\boldsymbol{I}(t) = (I_1(t), \ldots, I_D(t))$ 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.

$$\|w\|^2_{2,\rho\times\rho} := \int_{\mathcal{M}^2} \rho(\mathrm{d}z)\rho(\mathrm{d}z')w(z, z')^2 < \infty \tag{A.14}$$

yielding the general field equation:

$$\begin{aligned}\partial_t v_t(z) &= -\frac{1}{\tau}v_t(z) + \int_{\mathcal{M}} \rho(dz')w(z, z')\phi\left(v_t(z')\right) \\ &= -\frac{1}{\tau}v_t + K_w[\phi \circ v_t]\end{aligned} \tag{A.15}$$

where $K_w : \mathcal{H} \to \mathcal{H}$ is the Hilbert-Schmidt integral operator associated with the kernel w: [92]

$$K_w[\varphi](z) := \int_{\mathcal{M}} \rho(\mathrm{d}z')w(z, z')\varphi(z') \tag{A.16}$$

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

$$K_w[\varphi] = \sum_{\alpha\in\mathbb{N}} \lambda_\alpha x_\alpha \langle y_\alpha, \varphi\rangle_\rho \tag{A.17}$$

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. $\sum_{\alpha\in\mathbb{N}} \lambda_\alpha^2 = \|w\|^2_{2,\rho\times\rho} < \infty$ [92, 93]. The dynamics of general fields, Eq.(A.15), have thus a very similar structure to that of low-rank fields, Eq.(1.5) – 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:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\kappa_\alpha(t) &= -\frac{1}{\tau}\kappa_\alpha(t) + \lambda_\alpha\langle y_\alpha, \phi \circ v_t\rangle_\rho \qquad (\alpha \in I) \\ v_t &= \sum_{\alpha\in I} x_\alpha\kappa_\alpha(t) + e^{-t/\tau}v_0^{\perp}\end{aligned} \tag{A.18}$$

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.(A.17) [59], which corresponds to a model of low-rank network according to our framework (see Eq.(1.3)).

B Examples of collective dynamics

B.1 Fixed points in a generalised Hopfield model

We consider here the connectivity mappings given by Eq.(2.2), which read:

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

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.(1.9) write:

$$(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau})\boldsymbol{\kappa} = \int_{\mathbb{R}} Dz\, J\mathbf{A}\tilde{\boldsymbol{f}}(z)\phi(\boldsymbol{\kappa}\mathbf{A}\tilde{\boldsymbol{f}}(z)) \tag{B.2}$$

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:

$$\begin{aligned} (\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau})\boldsymbol{\lambda} &= \int_{\mathbb{R}} Dz\, J\mathbf{A}^T\mathbf{A}\tilde{\boldsymbol{f}}(z)\phi(\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{f}}(z)) \\ &= J\mathbf{A}^T\mathbf{A}\int_{\mathbb{R}} Dz \begin{pmatrix} [z]_+ \\ [z]_- \end{pmatrix} \phi(\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{f}}(z)) \end{aligned} \tag{B.3}$$

Notice that:

$$\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{f}}(z) = \lambda_1[z]_+ + \lambda_2[z]_- = \begin{cases} \lambda_1 z \ \ (z \geq 0) \\ \lambda_2 z \ \ (z < 0) \end{cases} \tag{B.4}$$

so that Eq.(B.3) can be written, by splitting the integral over $\mathbb{R}$:

$$(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau})\boldsymbol{\lambda} = J\mathbf{A}^T\mathbf{A}\left[\begin{pmatrix} 1 \\ 0 \end{pmatrix}\int_0^\infty Dz\, z\, \phi(\lambda_1 z) + \begin{pmatrix} 0 \\ 1 \end{pmatrix}\int_{-\infty}^0 Dz\, z\, \phi(\lambda_2 z)\right] \tag{B.5}$$

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

$$(\tau\frac{\mathrm{d}}{\mathrm{d}t} + 1)\boldsymbol{\lambda} = \tau JR\langle[z]_+\rangle\mathbf{A}^T\mathbf{A}\left[\begin{pmatrix} 1 \\ 0 \end{pmatrix}\mathbb{1}_{\{\lambda_1>0\}} - \begin{pmatrix} 0 \\ 1 \end{pmatrix}\mathbb{1}_{\{\lambda_2<0\}}\right] \tag{B.6}$$

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

$$\boldsymbol{\lambda}^{(+)} = k\mathbf{A}^T\mathbf{A}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = k\begin{pmatrix} 1 \\ \cos\alpha \end{pmatrix} \quad ; \quad \boldsymbol{\lambda}^{(-)} = -k\mathbf{A}^T\mathbf{A}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = k\begin{pmatrix} -\cos\alpha \\ -1 \end{pmatrix} \tag{B.7}$$

We can write:

$$
\begin{aligned}
(\tau\frac{\mathrm{d}}{\mathrm{d}t}+1)\boldsymbol{\lambda} &= \boldsymbol{\lambda}^{(+)}\mathbb{1}_{\{\lambda_1>0\}}+\boldsymbol{\lambda}^{(-)}\mathbb{1}_{\{\lambda_2<0\}} \\
&= \begin{cases} \boldsymbol{\lambda}^{(+)} & (\boldsymbol{\lambda}\in Q_1) \\ 0 & (\boldsymbol{\lambda}\in Q_2) \\ \boldsymbol{\lambda}^{(-)} & (\boldsymbol{\lambda}\in Q_3) \\ \boldsymbol{\lambda}^{(+)}+\boldsymbol{\lambda}^{(-)} & (\boldsymbol{\lambda}\in Q_4) \end{cases}
\end{aligned}
\tag{B.8}
$$

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:

$$
\boldsymbol{\kappa}^{(+)}=k\begin{pmatrix}1\\0\end{pmatrix} \quad ; \quad \boldsymbol{\kappa}^{(-)}=k\begin{pmatrix}-\cos\alpha\\+\sin\alpha\end{pmatrix}\approx k\begin{pmatrix}-1/2\\0.87\end{pmatrix}
\tag{B.9}
$$

and $(\boldsymbol{\kappa}^{(+)}+\boldsymbol{\kappa}^{(-)})$. The numerical values correspond to Fig.3(B), where $\alpha=\pi/3$ (i.e. $a=\cos\alpha=1/2$).

B.2 Different models of limit-cycle collective dynamics

Ring model.

For the ring model, we can write from Eq.(8):

$$
W_{ij}=J\boldsymbol{f}(z_i)\cdot\boldsymbol{g}(z_j), \quad \text{with:} \quad \boldsymbol{f}(z)=\begin{pmatrix}\cos z\\ \sin z\end{pmatrix} \quad , \quad \boldsymbol{g}=\boldsymbol{f}(z+\delta)
\tag{B.10}
$$

and the collective dynamics resulting from Eq.(1.9) write:

$$
(\frac{\mathrm{d}}{\mathrm{d}t}+\frac{1}{\tau})\boldsymbol{\kappa}=J\int_0^{2\pi}\frac{dz}{2\pi}\boldsymbol{g}(z)\phi(\boldsymbol{f}(z)\cdot\boldsymbol{\kappa})
\tag{B.11}
$$

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:

$$
(\frac{\mathrm{d}}{\mathrm{d}t}+\frac{1}{\tau})\boldsymbol{\kappa}=(\dot{\varrho}+\frac{1}{\tau}\varrho)\hat{e}_\varrho+\varrho\dot{\theta}\hat{e}_\theta \quad ; \quad \boldsymbol{f}(z)=\cos(z-\theta)\hat{e}_\varrho+\sin(z-\theta)\hat{e}_\theta
\tag{B.12}
$$

hence:

$$
\begin{aligned}
\begin{pmatrix}\dot{\varrho}+\frac{1}{\tau}\varrho\\ \varrho\dot{\theta}\end{pmatrix} &= J\int_0^{2\pi}\frac{dz}{2\pi}\begin{pmatrix}\cos(z+\delta-\theta)\\ \sin(z+\delta-\theta)\end{pmatrix}\phi(\varrho\cos(z-\theta)) \\
&= J\int_0^{2\pi}\frac{dz}{2\pi}\begin{pmatrix}\cos(z+\delta)\\ \sin(z+\delta)\end{pmatrix}\phi(\varrho\cos z)
\end{aligned}
\tag{B.13}
$$

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:

$$
\begin{pmatrix}\dot{\varrho}+\frac{1}{\tau}\varrho\\ \varrho\dot{\theta}\end{pmatrix}=\frac{JR}{2\pi}\int_{-\pi/2}^{\pi/2}dz\begin{pmatrix}\cos(z+\delta)\\ \sin(z+\delta)\end{pmatrix}=\frac{JR}{\pi}\begin{pmatrix}\cos\delta\\ \sin\delta\end{pmatrix}
\tag{B.14}
$$

which is identical to Eq.(2.5).

Gaussian distribution of neuronal locations.

Let the neurons be normally distributed in $\mathbb{R}^2$, i.e. $z_i \sim \mathcal{N}(0, \mathbb{1}_2)$. As discussed in Methods 2, this distribution can be parameterised in polar coordinates: each neuron is independently assigned a radius $\xi_i \geq 0$ and an angle $\varphi_i \sim \mathcal{U}([0, 2\pi[)$. The connectivity of Eq.(9) thus writes:

$$W_{ij} = \tilde{J} z_i R^{(\delta)} z_j = J' \begin{pmatrix} \xi_i \cos\varphi_i \\ \xi_i \sin\varphi_i \end{pmatrix} \begin{pmatrix} \xi_j \cos(\varphi_j + \delta) \\ \xi_j \sin(\varphi_j + \delta) \end{pmatrix} \tag{B.15}$$

which can be written, as a function of a neuron's radius ξ and angle φ:

$$W_{ij} = J' \tilde{\boldsymbol{f}}(z_i) \cdot \tilde{\boldsymbol{g}}(z_j) \quad \text{with:} \quad \tilde{\boldsymbol{f}}(\xi, \varphi) = \begin{pmatrix} \xi \cos\varphi \\ \xi \sin\varphi \end{pmatrix} = \xi \boldsymbol{f}(\varphi) \quad ; \quad \tilde{\boldsymbol{g}}(\xi, \varphi) = \xi \boldsymbol{g}(\varphi) \tag{B.16}$$

where $\boldsymbol{f}, \boldsymbol{g}$ are defined by Eq.(B.10). The collective dynamics of Eq.(1.9) thus write:

$$(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau})\boldsymbol{\kappa} = J' \int_0^{2\pi} \frac{d\varphi}{2\pi} \int_{\mathbb{R}^+} \rho_\xi(d\xi)\ \xi \boldsymbol{g}(\varphi) \phi(\xi \boldsymbol{f}(\varphi) \cdot \boldsymbol{\kappa}) \tag{B.17}$$

where $\rho_\xi(d\xi)$ denotes the distribution of the radii (it is the χ_2 distribution if neuronal locations are normally distributed in the plane). In the case of a step transfer function ϕ, we have $\phi(\xi x) = \phi(x)\ \forall \xi > 0$, hence:

$$(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau})\boldsymbol{\kappa} = J' \Big(\int_{\mathbb{R}^+} \xi \rho_\xi(d\xi) \Big) \int_0^{2\pi} \frac{d\varphi}{2\pi}\ \boldsymbol{g}(\varphi) \phi(\boldsymbol{f}(\varphi) \cdot \boldsymbol{\kappa}) \tag{B.18}$$

which is identical to Eq.(B.11) if $J' \langle \xi \rangle = J$, where $\langle \xi \rangle := \int \xi \rho_\xi(d\xi)$ ($= \sqrt{\pi/2}$ for the χ_2-distribution). 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 $\langle \xi \rangle$.

Arbitrary transfer function and radial distribution

Here, we extend the above derivation of the collective dynamics for the connectivity given by Eq.(B.15) to the case of an arbitrary transfer function ϕ and rotationally invariant distribution of the neuronal locations. In this case, the collective dynamics (Eq.(1.9)) corresponding to the connectivity functions of Eq.(B.16) are:

$$(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau}\boldsymbol{\kappa}) = J \int_0^{2\pi} \frac{d\varphi}{2\pi} \int_{\mathbb{R}^+} \rho_\xi(d\xi)\ \xi \boldsymbol{f}(\varphi + \delta) \phi(\xi \boldsymbol{f}(\varphi) \cdot \boldsymbol{\kappa}) \tag{B.19}$$

where $\boldsymbol{f} = (\cos, \sin)$ (Eq.(B.10)), and $\rho_\xi(d\xi)$ is the distribution of the radii of neuronal locations. For the radial coordinate (ϱ, θ) of $\boldsymbol{\kappa}$, it writes:

$$\begin{aligned}\begin{pmatrix}\dot{\varrho} + \frac{1}{\tau}\varrho \\ \varrho\dot{\theta}\end{pmatrix} &= J\int_0^{2\pi}\frac{d\varphi}{2\pi}\int_{\mathbb{R}^+}\rho_\xi(d\xi)\ \xi\begin{pmatrix}\cos(\varphi+\delta-\theta)\\ \sin(\varphi+\delta-\theta)\end{pmatrix}\phi(\xi\varrho\cos(\varphi-\theta)) \\ &= J\int_0^{2\pi}\frac{d\varphi}{2\pi}\int_{\mathbb{R}^+}\rho_\xi(d\xi)\ \xi\begin{pmatrix}\cos(\varphi+\delta)\\ \sin(\varphi+\delta)\end{pmatrix}\phi(\xi\varrho\cos(\varphi))\end{aligned} \tag{B.20}$$

Using the sum expansion of the cos and sin functions, we get for the first component of Eq.(B.20):

$$\int_0^{2\pi}\frac{d\varphi}{2\pi}\cos(\varphi+\delta)\phi(\xi\varrho\cos(\varphi)) = \cos\delta\int_0^{2\pi}\frac{d\varphi}{2\pi}\cos\varphi\phi(\xi\varrho\cos(\varphi)) - \sin\delta\underbrace{\int_0^{2\pi}\frac{d\varphi}{2\pi}\sin\varphi\phi(\xi\varrho\cos(\varphi))}_{=0} \tag{B.21}$$

where the last term is zero because the integrand is odd; and similarly, for the second component:

$$\int_0^{2\pi}\frac{d\varphi}{2\pi}\sin(\varphi+\delta)\phi(\xi\varrho\cos(\varphi)) = \sin\delta\int_0^{2\pi}\frac{d\varphi}{2\pi}\cos(\varphi)\phi(\xi\varrho\cos(\varphi)) \tag{B.22}$$

Thus, we obtain from Eq.(B.20):

$$\begin{pmatrix}\dot{\varrho} + \frac{1}{\tau}\varrho \\ \varrho\dot{\theta}\end{pmatrix} = J\begin{pmatrix}\cos\delta\\ \sin\delta\end{pmatrix}\cdot\Phi(\varrho) \tag{B.23}$$

where $\Phi(\varrho)$ is defined by (after the change of variable $x = \cos\varphi$):

$$\Phi(\varrho) := \int_{\mathbb{R}^+}\rho_\xi(d\xi)\int_0^{2\pi}\frac{d\varphi}{2\pi}\xi\cos\varphi\,\phi(\varrho\xi\cos\varphi) = \int_{\mathbb{R}^+}\rho_\xi(d\xi)\int_{-1}^{1}\frac{dx}{\pi\sqrt{1-x^2}}\xi x\,\phi(\varrho\xi x) \tag{B.24}$$

Provided that the fixed point equation for ϱ,

$$\varrho = \tau J\cos(\delta)\,\Phi(\varrho), \tag{B.25}$$

admits a solution ϱ^*, the model generates rotational dynamics that converge towards a limit cycle of radius ϱ^*.

Noticeably, the angular speed ω of the limit cycle does not depend on the transfer function ϕ, nor on the distribution ρ_ξ of the radii, but only on the parameters τ and δ. From Eq.(B.23) and using the definition of ϱ^* (Eq.(B.25)), we get

$$\omega = \dot{\theta}|_{\varrho=\varrho^*} = J\sin\delta\Phi(\varrho^*)/\varrho^* = \frac{1}{\tau}\tan\delta. \tag{B.26}$$

C Symmetries

We show here the results of Sec.2, relating the symmetries of a field model to those of its collective dynamics. We take an operator theory perspective. The dynamics of the field, Eq.(1.6), can be

written:

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

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 $\mathcal{F} = \mathrm{span}\{f_\mu\}$) is a nonlinear operator defined by:

$$\mathcal{D}_{\mathcal{H}} : \varphi \mapsto \sum_\mu f_\mu \langle g_\mu, \phi \circ \varphi \rangle_\rho \tag{C.2}$$

On the other hand, the collective dynamics of Eq.(1.9) are described by a nonlinear operator $\mathcal{D}_D$ acting on $\mathbb{R}^D$:

$$\begin{aligned} &\Box \boldsymbol{\kappa} = \mathcal{D}_D \boldsymbol{\kappa} \\ \text{with} \quad &\mathcal{D}_D : \boldsymbol{\kappa} \mapsto \langle \boldsymbol{g}, \phi \circ (\boldsymbol{f} \cdot \boldsymbol{\kappa}) \rangle_\rho \end{aligned} \tag{C.3}$$

Defining symmetries

In the following, let $\mathcal{S}$ be the set of bijections $S : \mathcal{M} \to \mathcal{M}$ such that $\varphi \circ S \in \mathcal{H}$, $\forall \varphi \in \mathcal{H}$. For any mapping $S \in \mathcal{S}$, the composition operator $L^{(S)} : \varphi \mapsto \varphi \circ S$ defines a linear operator on $\mathcal{H}$.

We call a *(microscopic) symmetry* of a field model a bijection $S \in \mathcal{S}$ such that the operator $\mathcal{D}_{\mathcal{H}}$ describing the field 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.(C.1), then $L^{(S)} v_t$ is also a solution to the dynamics, with initial condition $L^{(S)} v_0$. Analogously, a *macroscopic* symmetry is a symmetry of the collective dynamics; i.e., a mapping $T : \mathbb{R}^D \to \mathbb{R}^D$ that commutes with $\mathcal{D}_D$.

Results

We show that any microscopic symmetry $S \in \mathcal{S}$ of the field model generates a macroscopic symmetry $T^{(S)}$ of its collective dynamics, for which we give an explicit expression – under an additional condition that is satisfied by invariance symmetries.

Theorem 1. *Let $S \in \mathcal{S}$.*

1. If (i) S is a symmetry of the field model, and (ii) $\mathcal{F} = \mathrm{span}\{f_\mu\}$ 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

$$T^{(S)} = M(\langle f_\mu, f_\nu \circ S \rangle_\rho) \tag{C.4}$$

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 conditions (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 $(g, z) \mapsto S_g(z)$ denote the action of a group G on $\mathcal{M}$. It is called a *left* group action if $S_{gh} = S_g \circ S_h$, i.e. $S_{gh}(z) = S_g(S_h(z))$, $\forall g, h \in G$ (and a *right* group action if $S_{gh} = S_h \circ S_g$). G is a *symmetry group* of the field model if for any $g \in G$, S_g is a symmetry. It is a *left* symmetry group if it has a left group action on $\mathcal{M}$ (and respectively, a *right* symmetry group if it has a right group action).

Similarly, a left (or right) symmetry group of the collective dynamics is a group with a left (resp., right) action 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, such that the action S_g of each element $g \in G$ 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.

Let $S \in \mathcal{S}$; and $L^{(S)} : \mathcal{H} \to \mathcal{H} : \varphi \mapsto \varphi \circ S$ be the linear composition operator on $\mathcal{H}$.

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 $V : \mathbb{R}^D \to \mathcal{F} \subset \mathcal{H}$:

$$K : \varphi \mapsto M\langle \boldsymbol{f}, \varphi \rangle_\rho \quad ; \quad V : \boldsymbol{k} \mapsto \boldsymbol{k} \cdot \boldsymbol{f} \tag{C.5}$$

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

$$KV = \mathbb{1}_{\mathbb{R}^D} \quad ; \quad VK = \Pi_{\mathcal{F}} \tag{C.6}$$

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.(C.1) the following expression for the dynamics of $\boldsymbol{\kappa}$:

$$\Box\boldsymbol{\kappa} = \Box K v_t = K \Box v_t = K\mathcal{D}_{\mathcal{H}} v_t = K\mathcal{D}_{\mathcal{H}} V \boldsymbol{\kappa}(t) \tag{C.7}$$

which can be identified with Eq.(C.3). □

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 $\varphi, \psi \in \mathcal{H}$,

$$\begin{aligned} \langle \varphi, L^{(S)}\psi \rangle_\rho &= \langle \varphi, \psi \circ S \rangle_\rho = \int_{\mathcal{M}} \varphi(z)\psi(Sz)\rho(\mathrm{d}z) \\ &\overset{(*)}{=} \int_{\mathcal{M}} \varphi(S^{-1}z)\psi(z)\,(\rho \circ S^{-1})(\mathrm{d}z) = \int_{\mathcal{M}} \varphi(S^{-1}z)\psi(z)\,\rho(\mathrm{d}z) \\ &= \langle \varphi \circ S^{-1}, \psi \rangle_\rho = \langle L^{(S^{-1})}\varphi, \psi \rangle_\rho \end{aligned} \tag{C.8}$$

where we used a change of variable in $(*)$. □

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):

$$K(L\mathcal{D}_{\mathcal{H}})V = K(\mathcal{D}_{\mathcal{H}}L)V \tag{C.9}$$

The left-hand side can be written:

$$KL\mathcal{D}_{\mathcal{H}}V = KL(VK\mathcal{D}_{\mathcal{H}})V = (KLV)(K\mathcal{D}_{\mathcal{H}}V) = T\mathcal{D}_D \tag{C.10}$$

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.(C.6); and we used Lemma 1 in the last equality.

In turn, the right-hand side writes:

$$K\mathcal{D}_{\mathcal{H}}(LV) = K\mathcal{D}_{\mathcal{H}}(VKLV) = (K\mathcal{D}_{\mathcal{H}}V)(KLV) = \mathcal{D}_D T \tag{C.11}$$

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 $\boldsymbol{\kappa} \in \mathbb{R}^D$:

$$\begin{aligned}(T\boldsymbol{\kappa})_\mu &= \sum_\nu M_{\mu\nu}\langle f_\nu, L^{(S)}(\sum_\sigma f_\sigma\kappa_\sigma)\rangle_\rho = \sum_\nu M_{\mu\nu}\langle f_\nu, \sum_\sigma \kappa_\sigma(L^{(S)}f_\sigma)\rangle_\rho \\ &= \sum_\sigma\sum_\nu M_{\mu\nu}\langle f_\nu, f_\sigma\circ S\rangle_\rho\kappa_\sigma =: \sum_\sigma T^{(S)}_{\mu\sigma}\,\kappa_\sigma\end{aligned} \tag{C.12}$$

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:

$$\begin{aligned}(\mathcal{D}_{\mathcal{H}}L^{(S)}v)(z) &= \sum_\mu f_\mu(z)\langle g_\mu, \phi\circ(L^{(S)}v)\rangle_\rho \\ &= \sum_\mu f_\mu(z)\langle g_\mu, \phi\circ v\circ S\rangle_\rho = \sum_\mu f_\mu(z)\langle g_\mu, L^{(S)}(\phi\circ v)\rangle_\rho \\ &= \sum_\mu f_\mu(z)\langle {L^{(S)}}^\dagger g_\mu, \phi\circ v\rangle_\rho \overset{(*)}{=} \sum_\mu f_\mu(z)\langle g_\mu\circ S^{-1}, \phi\circ v\rangle_\rho \\ &= \int_{\mathcal{M}} w(z,S^{-1}z')\phi\circ v(z)\rho(\mathrm{d}z) \overset{(**)}{=} \int_{\mathcal{M}} w(Sz,z')\phi\circ v(z)\rho(\mathrm{d}z) \\ &= \sum_\mu f_\mu(Sz)\langle g_\mu, \phi\circ v\rangle_\rho = \sum_\mu (L^{(S)}f_\mu)(z)\langle g_\mu, \phi\circ v\rangle_\rho \\ &= (L^{(S)}\mathcal{D}_{\mathcal{H}}v)(z)\end{aligned} \tag{C.13}$$

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:

$$\begin{aligned} w \text{ invariant for } S \ &\Leftrightarrow\ \sum_\mu f_\mu(Sz) g_\mu(Sz') = \sum_\mu f_\mu(z) g_\mu(z') \quad \forall z, z' \in \mathcal{M} \\ &\Leftrightarrow\ \sum_\mu (L^{(S)} f_\mu)(z)(L^{(S)} g_\mu)(z') = \sum_\mu f_\mu(z) g_\mu(z') \quad \forall z, z' \\ &\overset{(*)}{\Rightarrow}\ \sum_\mu (L^{(S)} f_\mu)(z) \langle L^{(S)} g_\nu, L^{(S)} g_\mu \rangle_\rho = \sum_\mu f_\mu(z) \langle L^{(S)} g_\nu, g_\mu \rangle_\rho \quad \forall z, \forall \nu \\ &\Rightarrow\ \sum_\mu G_{\nu\mu} L^{(S)} f_\mu = \sum_\mu f_\mu \langle L^{(S)} g_\nu, g_\mu \rangle_\rho \in \mathcal{F} \quad \forall \nu = 1, \dots, D \end{aligned} \tag{C.14}$$

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.(C.14) yields:

$$L^{(S)} f_\mu = \sum_\nu (G^{-1})_{\mu\nu} \sum_\sigma \langle L^{(S)} g_\nu, g_\sigma \rangle_\rho \, f_\sigma \in \mathcal{F} \quad \forall \mu = 1, \dots, D \tag{C.15}$$

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 G has 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:

$$T^{(gh)} = K L^{(gh)} V = K L^{(h)} L^{(g)} V = K L^{(h)} (VK) L^{(g)} V = T^{(h)} T^{(g)} \tag{C.16}$$

where we used Lemma 2 in the third equality, with $VK = \Pi_{\mathcal{F}}$ (Eq.(C.6)).

Notice eventually that the action on $\mathbb{R}^D$ of the identity element $e \in G$ is the identity: 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.(C.6).

D Adding heterogeneity in a field model

Rescaling of outgoing weights

Consider a field model, obeying the collective dynamics of Eq.(1.9), 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 an arbitrary distribution ρ_ξ. This results in a "new" field model, where each neuron

i is also characterised by its heterogeneity parameter ξ_i, corresponding to an additional dimension of the similarity space. We denote each neuron's location in the "new" model

$$\tilde{z}_i = (z_i, \xi_i) \in \mathcal{M} \times \mathbb{R}_+ \tag{D.1}$$

and the "new" connectivity reads:

$$\tilde{W}_{ij} = \tilde{\boldsymbol{f}}(\tilde{z}_i) \cdot \tilde{\boldsymbol{g}}(\tilde{z}_j) = \xi_j W_{ij} = \boldsymbol{f}(z_i) \cdot \xi_j \boldsymbol{g}(z_j) \tag{D.2}$$

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.(1.9), of the latent variables $\tilde{\boldsymbol{\kappa}}$ associated with this "new" model with heterogeneous connectivity write:

$$(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau})\tilde{\boldsymbol{\kappa}} = \langle \xi \rangle \cdot \langle \boldsymbol{g}, \phi \circ (\boldsymbol{f} \cdot \tilde{\boldsymbol{\kappa}}) \rangle_\rho \tag{D.3}$$

with $\langle \xi \rangle := \int \xi \rho_\xi(d\xi)$ the mean value of the heterogeneity parameter ξ. This is identical to the dynamics of the original model (Eq.(1.9)), up to a rescaling of the latent variables by a factor $\langle \xi \rangle$.

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" location $\tilde{z}_i = (z_i, a_i)$: we write $\tilde{\phi}(h, \tilde{z}_i) = a_i \phi(h)$. Similar to the example above ("Rescaling of outgoing 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, $\langle a \rangle = \int a \rho_a(da)$.

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 [49, 60]. 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:

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

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 location $z_i \in \mathbb{R}^{2D}$ 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 [60].

E.2 Field models over one-dimensional space

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

$$\hat{D}(\boldsymbol{\kappa}) = \int_0^1 g(z)\phi\Big(\sum_{\nu=1}^{D} \kappa_\nu f_\nu(z) + f_0(z)\Big)\mathrm{d}z \qquad \text{with } g, \{f_\mu\}_{0\le\mu\le D} \subset \mathcal{H}_0 \tag{E.2}$$

AF_0 is a subset of the linear readouts that can be generated by field models over the one-dimensional similarity space $\mathcal{M} = [0, 1]$.

Consider an arbitrary dynamical system with asymptotic exponential decay:

$$\frac{\mathrm{d}}{\mathrm{d}t}\kappa_\mu(t) = -\frac{1}{\tau}\kappa_\mu(t) + D_\mu(\vec{\kappa}(t)) \quad (\alpha = 1, \dots, D) \tag{E.3}$$

where $D_\mu \in \mathcal{C}(\mathbb{R}^D, \mathbb{R})$ 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 feed-forward neural networks, this can be decomposed into two parts [94]. 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:

$$\begin{aligned}
(c_1\hat{D}^{(1)} + c_2\hat{D}^{(2)})(\vec{\kappa}) &= c_1 \int_0^1 g^{(1)}(z)\phi\left(\vec{\kappa}\cdot\boldsymbol{f}^{(1)}(z) + f_0^{(1)}(z)\right)\mathrm{d}z \\
&\quad + c_2 \int_0^1 g^{(2)}(z)\phi\left(\vec{\kappa}\cdot\boldsymbol{f}^{(2)}(z) + f_0^{(2)}(z)\right)\mathrm{d}z \\
&= c_1 \int_0^{1/2} 2g^{(1)}(2z)\phi\left(\vec{\kappa}\cdot\boldsymbol{f}^{(1)}(2z) + f_0^{(1)}(2z)\right)\mathrm{d}z \\
&\quad + c_2 \int_{1/2}^1 2g^{(2)}(2z-1)\phi\left(\vec{\kappa}\cdot\boldsymbol{f}^{(2)}(2z-1) + f_0^{(2)}(2z-1)\right)\mathrm{d}z \\
&= \int_0^1 g'(z)\phi\left(\vec{\kappa}\cdot\vec{f}'(z) + f_0'(z)\right)\mathrm{d}z \quad =: \hat{D}'(\vec{\kappa})
\end{aligned} \tag{E.4}$$

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

$$g'(z) = \begin{cases} 2c_1 g^{(1)}(2z) & \text{if } z \in [0, 1/2[\\ 2c_2 g^{(2)}(2z-1) & \text{if } z \in [1/2, 1] \end{cases} \; ; \; \boldsymbol{f}'(z) = \begin{cases} \boldsymbol{f}^{(1)}(2z) & \text{if } z \in [0, 1/2[\\ \boldsymbol{f}^{(2)}(2z-1) & \text{if } z \in [1/2, 1] \end{cases} \tag{E.5}$$

for $\alpha = 0, \ldots, D$. Because of the vanishing boundary condition on $[0, 1]$, if $g^{(1)}, \{f_\mu^{(1)}\}$, $g^{(2)}, \{f_\mu^{(2)}\}$ are all in $\mathcal{H}_0$, so are $g', \{f'_\mu\}$, and thus $\hat{D}' \in \mathrm{AF}_0$. Therefore, AF_0 is a vector subspace of $\mathcal{C}(\mathbb{R}^D, \mathbb{R})$.

Elementary functions.

Following the line of [94], 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 $\boldsymbol{f}(z) = AT(z)\boldsymbol{d}$, $f_0(z) = -AT(z)b$ and $g(z) = T(z)$ in Eq.(E.2). Then for $A \to \infty$,

$$\hat{D}_A(\boldsymbol{\kappa}) := \int_0^1 T(z)\phi\left(AT(z)\big(\boldsymbol{\kappa} \cdot \boldsymbol{d} - b\big)\right) \mathrm{d}z \to \begin{cases} \phi(+\infty) & (\boldsymbol{\kappa} \cdot \boldsymbol{d} > b) \\ \phi(-\infty) & (\boldsymbol{\kappa} \cdot \boldsymbol{d} < b) \\ \phi(0) & (\boldsymbol{\kappa} \cdot \boldsymbol{d} = b) \end{cases} \tag{E.6}$$

By assuming that ϕ is sigmoidal (i.e. $\phi(+\infty) = 1, \phi(-\infty) = 0$), we see that $\hat{D}_A \in \mathrm{AF}_0$ 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 [94], arbitrary functions D_α can be approximated by functions of the form of Eq.(E.2).