nature reviews neuroscience

https://doi.org/10.1038/s41583-023-00693-x

Perspective

Check for updates

A unifying perspective on neural manifolds and circuits for cognition

Christopher Langdon[1,2,3], Mikhail Genkin[2,3] & Tatiana A. Engel[1,2] ✉

Abstract

Two different perspectives have informed efforts to explain the link between the brain and behaviour. One approach seeks to identify neural circuit elements that carry out specific functions, emphasizing connectivity between neurons as a substrate for neural computations. Another approach centres on neural manifolds – low-dimensional representations of behavioural signals in neural population activity – and suggests that neural computations are realized by emergent dynamics. Although manifolds reveal an interpretable structure in heterogeneous neuronal activity, finding the corresponding structure in connectivity remains a challenge. We highlight examples in which establishing the correspondence between low-dimensional activity and connectivity has been possible, unifying the neural manifold and circuit perspectives. This relationship is conspicuous in systems in which the geometry of neural responses mirrors their spatial layout in the brain, such as the fly navigational system. Furthermore, we describe evidence that, in systems in which neural responses are heterogeneous, the circuit comprises interactions between activity patterns on the manifold via low-rank connectivity. We suggest that unifying the manifold and circuit approaches is important if we are to be able to causally test theories about the neural computations that underlie behaviour.

Sections

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA. [2]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [3]These authors contributed equally: Christopher Langdon and Mikhail Genkin. ✉e-mail: tatiana.engel@princeton.edu

Introduction

Behavioural and cognitive functions emerge from the dynamic interactions of many neurons wired into circuits. Understanding how circuit connectivity gives rise to neural dynamics and behaviour is a central goal in systems neuroscience. This problem, however, remains unresolved, in part owing to the experimental challenge of measuring both the connectivity and the activity of the same neurons during behaviour. Although stunning technological advances have enabled us to record activity from increasingly large populations of neurons[1–3], the observational data that these experiments generate do not unambiguously point to circuit mechanisms. Likewise, reconstructions of anatomical connectivity[4–7] constrain the space of possible neural dynamics but do not uniquely predict the activity patterns that arise from the circuit to control specific behaviours. Therefore, theory and computational modelling have been instrumental in bridging the gaps between circuit connectivity, neural dynamics and behavioural functions.

Traditionally, theoretical models hypothesize possible circuit mechanisms to reproduce the neural responses and behaviour observed in experiments[8–14]. In these circuit models, the recurrent connectivity is usually hand-crafted to produce the neural activity patterns that are needed to solve a particular behavioural task. For example, clustered connectivity (in which there are stronger connections within than between clusters of neurons) gives rise to discrete attractors (self-sustained and stable states of the system) in neural dynamics, which can support categorical decision-making[10,15–17]. When triggered by sensory input, the network activity converges to one of the attractors, each of which represents a different choice alternative[10,18]. In a similar manner, circuit models have been used to relate connectivity structure to the dynamical-system description of neural computation across many cognitive tasks[19,20]. Such links are powerful because they enable us to predict the behavioural effects of circuit perturbations (such as changes in the excitation–inhibition balance[21,22]), opening up the possibility that we can experimentally test the hypothesized causal mechanisms[23–25].

However, recently available large-scale recordings have exposed a rich complexity of neural responses in the brain[26] that cannot be explained by classical circuit models, which usually assume a simple hand-crafted connectivity structure and, thus, produce functionally homogeneous neural responses. For example, in the discrete attractor network model with clustered connectivity described above, all neurons within a single cluster show the same selectivity for one choice and respond with a similar time course[10,18]. By contrast, recordings from cortical neurons during cognitive tasks show that single neurons exhibit complex mixed selectivity for multiple task variables and diverse temporal response profiles[27–30]. Tying this complex and heterogeneous activity to the underlying circuit mechanism thus poses a formidable challenge.

Over the past decade, multiple statistical techniques have emerged that can find structure in heterogeneous neural responses[31–34] (Box 1). Although diverse, the responses of different neurons in a population are usually tightly correlated during behavioural tasks, meaning that the population expresses only a restricted set of activity patterns. Geometrically, we can picture this set of permissible activity patterns as a surface in a neural population state space in which each axis represents the activity of one neuron (Box 1). This surface — referred to as the neural manifold — is often low-dimensional and reveals interpretable structure in the neural population activity related to behavioural task execution. Modelling how neural population dynamics unfold along the manifold as the task progresses provides a dynamical-system description of neural computation[35–39]. The discovery of interpretable manifolds in multiple brain areas and behavioural tasks suggests that computation through dynamics on a manifold may be a general principle for the organization of heterogeneous neural responses in the brain[40,41].

Although neural manifolds concisely summarize heterogeneous single-cell responses, they provide only a descriptive model of neural computation. Without links to causal mechanisms, the manifold description lacks the power to generate testable predictions for experiments. It is evident that correlations in neural responses arise from constraints posed by the underlying network connectivity; however, the relationship between the neural manifold structure and the connectivity that gave rise to it remains largely unappreciated.

In this Perspective, we synthesize recent theoretical and experimental work that links neural manifolds to their underlying circuit mechanisms, suggesting that the manifold and circuit perspectives on neural computation are inseparable. Although several recent reviews have highlighted insights provided by studies focusing on neural manifolds[32,40–42] and circuits[43,44] separately, we advocate here for the integration of neural manifold and circuit approaches to cognition. We review the fly's head direction system as an example of convergence between the manifold and circuit structure that has been confirmed experimentally. We then discuss recent theoretical work suggesting that similar convergence may exist in systems with distributed mixed selectivity. Experimental validation of this correspondence will require the connectivity and activity of the same neurons to be mapped or the model predictions to be tested in perturbation experiments. We argue that theorists and experimentalists should not satisfy themselves with descriptions of neural computations as dynamics on manifolds but should, instead, seek understanding that integrates circuit connectivity, dynamics and behaviour.

Circuit–manifold convergence: head direction system

The head direction system is the best-studied example of a convergence between a neural manifold and circuit structure that has been confirmed experimentally at the single-cell level[45–47]. The function of the head direction system is to represent the direction in which an animal is heading and to update this representation according to information received about the animal's angular velocity input and the position of visual landmarks.

The ring manifold for head direction

The head direction angle is a one-dimensional circular variable (Box 1), which is topologically equivalent to a ring (Fig. 1a). Consistent with this topology, neural responses in the head direction systems of the mouse[48] and the fruit fly *Drosophila melanogaster*[49] organize on a manifold with ring topology, such that the position of the neural population activity on the ring manifold parametrically encodes the head direction (Fig. 1a).

The ring manifold in the fly is beautifully conspicuous owing to the simple physical layout of the ellipsoid body, the core neuropil in the fly head direction system[49], within which topographically organized neural responses can be directly visualized using calcium imaging (Fig. 1b). The ellipsoid body has a circular structure and the neurons within the ellipsoid body (called E-PG neurons) display a 'bump' of activity within this circle, indicating that a small set of neighbouring neurons is active, at any given time. The location of this activity bump within the circle precisely tracks the actual head direction of the fly (Fig. 1b), moving as it rotates[49,50] and adjusting relative to visual landmarks[49]. Furthermore, only one activity bump is present at any given time and an existing activity bump disappears when another artificial bump is induced by optogenetic stimulation[51,52]. Together, the uniqueness of the bump and its movement with the angular velocity establish the functional significance of the ring manifold.

By contrast, the rodent's head direction system is more complex, involving many brain regions in which head direction cells are scattered, and there is no evidence so far for a topographic ring structure organization[46,53]. Nevertheless, the responses of head direction cells in the anterodorsal thalamic nucleus[48] and post-subiculum[54] of mice also form a ring manifold. The ring is nonlinearly embedded, twisting through various dimensions in the high-dimensional neural population space[48] (Fig. 1a). The manifold embedding is jointly defined by the tuning curves of all neurons in the population[34] (Box 1). Some of these neurons have complex, multimodal tuning to the head direction[48,54], which can affect the manifold embedding without changing the underlying ring topology. Thus, complex tuning can be consistent with a simple topology of neural population responses, and manifold analyses can reveal this simple structure in heterogeneous single-cell responses.

Ring attractor models

How does the ring manifold for head direction arise from circuit connectivity? Theoretical models can suggest an answer to this question. In one type of classical neural circuit model (Box 2), called a ring attractor model, the head direction cells are arranged on a circle and wired with strong local excitatory and uniform inhibitory connections[8,55,56] (Fig. 1c). Each cell has bell-shaped tuning for its preferred head orientation angle and, owing to the balance of excitation and inhibition, the network activity localizes into a single bump representing the current head orientation angle. In the absence of external inputs (such as those representing angular velocity or visual landmarks), the activity bump is persistent because of local recurrent excitation and unique owing to global inhibition (that is, only a single bump exists at all times). This ring attractor model also provides a circuit mechanism through which the head direction encoding on the manifold can be updated. Landmark cells are hypothesized to carry information about visual cues and to provide direct localized input to the corresponding head direction cells[57] (Fig. 1c), whereas clockwise and anticlockwise rotation cells are hypothesized to update the bump location in response to self-motion[8,58]. The rotation cells are activated by the angular velocity input and update the bump location via local asymmetric recurrent connections with the head direction cells. Each group of rotation cells receives input from its corresponding group of head direction cells and projects to the clockwise or anticlockwise neighbours of that head direction cell (Fig. 1c).

The ring attractor model accounts for all salient features of the ring manifold and dynamics in the fly ellipsoid body, although it does not reproduce the complex tuning curves of some head direction cells in rodents[48,54]. It is important to note that the topology of the connectivity can be distinct from its spatial layout in the brain tissue. That is, the ring topology in connectivity can arise in model networks in which neurons are not spatially arranged on a ring[59], consistent with the lack of spatial topography in the rodent head direction system[53]. Furthermore, very similar dynamics on a ring manifold can arise from different biophysical mechanisms. For example, modelling has shown that the ring manifold topology and dynamics can emerge from structured inhibition between rotation and head direction cells rather than local recurrent excitation[46,58], consistent with the lack of strong recurrent connections between head direction cells that has been observed in

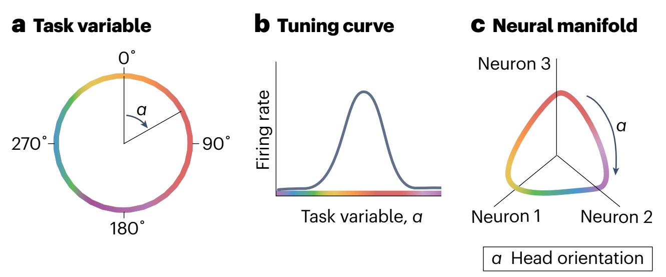
Box 1

Neural manifolds

Task variables refer to the discrete or continuous parameters of a task, as well as to related variables for intermediate representations and computations. These include stimulus parameters (such as the orientation or colour of a visual stimulus), variables that reflect the state of the environment (such as current spatial position or head orientation; see the figure, part **a**) and unobserved cognitive representations such as decision variables or accumulated evidence towards a choice. Unobserved variables are called latent variables.

Task variables are encoded by neural activity in the brain. The neural code can be summarized on the level of individual neurons in the form of tuning curves that describe a neuron's firing rate as a function of particular task variables (see the figure, part **b**). For example, neurons in the primary visual cortex have bell-shaped tuning curves for the orientation of a stimulus, with each curve being centred at the preferred orientation angle of the neuron[141]. In many brain regions, the responses of single neurons are complex and heterogeneous, with each neuron being tuned for a mixture of task variables[27–30,48] and many neurons responding to each task variable. These properties of the neural code are referred to as distributed mixed selectivity.

The joint activity of all neurons in a population can be described in a neural population state space, an *N*-dimensional Euclidian space in which each axis corresponds to the firing rate of one neuron (see the figure, part **c**). A population state is a point in this *N*-dimensional space and the evolution of neural responses over time forms a trajectory that reflects the collective dynamics of all neurons, called neural population dynamics[40]. In many tasks and brain areas, neural population activity does not explore all possible states but stays within a confined region of the state space[31–34,42]. The neural manifold is the continuous set of points in a state space that are explored by neural population activity (see the figure, part **c**). The position of neural population activity on a manifold at a given time often encodes task variables. For example, the ring manifold is a one-dimensional manifold parameterized by an angle α, thus the ring manifold can encode circular variables such as head direction[48] or stimulus location at a particular spatial angle in a working memory task[29]. In some cases, the neural manifold can be nonlinearly embedded in the high-dimensional neural population state space. The shape of this embedding (that is, the manifold's geometry) depends on the tuning curves of all neurons in the population[34].



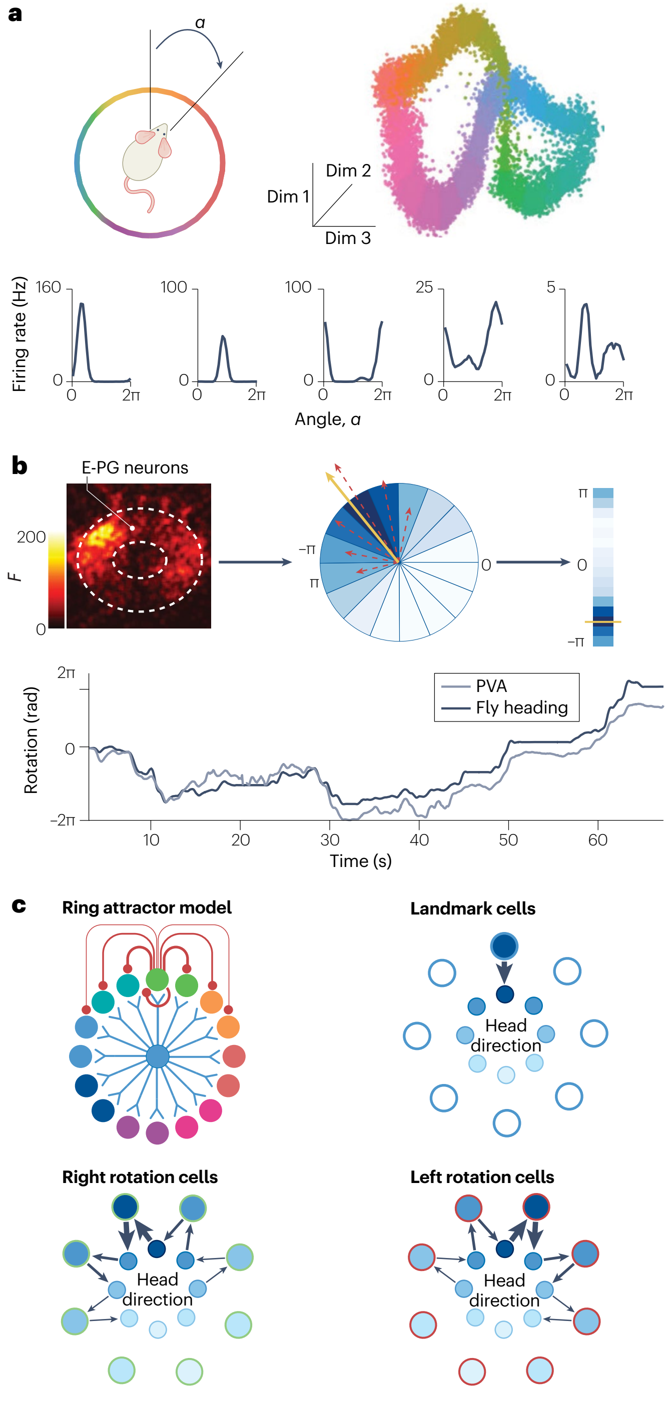


Fig. 1 | Convergence of a manifold and circuit in the head direction system. **a**, Head direction angle is a one-dimensional circular variable, *α* (upper left). A neural manifold discovered from neural population activity recorded in the anterodorsal thalamic nucleus of mice as they explore their surroundings takes the form of a one-dimensional ring that smoothly encodes head direction[48]. This is shown in a visualization of the manifold created using the Isomap algorithm[140] (upper right). Each dot corresponds to the population activity state at a single time. Colour coding represents parameterization of the manifold by a one-dimensional circular variable *α*, which closely matches the measured heading angle up to a choice of origin and direction. The ring manifold is nonlinearly embedded in the neural population state space. The shape of this embedding is determined by the heterogeneous and nonlinear tuning curves of individual neurons, five examples of which are shown in the lower panel[48]. **b**, E-PG neurons in the *Drosophila melanogaster* head direction system are arranged in a circle within the ellipsoid body of the fly brain. Calcium imaging reveals a localized 'bump' of neuronal activity within the ring encoding the head direction at any given time (upper left; *F*, fluorescence intensity in arbitrary units)[49,50]. Upper centre: the population vector average (PVA; gold arrow) estimates the position of the centre of the bump on the circle by summing vectors (dashed red arrows) pointing in the directions of each of 22.5° wedges around the ellipsoid body with length equal to the instantaneous calcium activity in each wedge. Shade of blue indicates calcium activity in each wedge. The ellipsoid body is unwrapped in a vertical axis (upper right; gold bar indicates angle of PVA) to display the population time series in the lower panel. The PVA accurately tracks the actual head direction of the animal (lower panel)[50]. **c**, A ring attractor model accounts for the ring manifold topology and dynamics of the head direction cells[55,56]. The model consists of head direction cells arranged in a ring and receiving local excitatory (red) and uniform inhibitory (blue) connections (upper left; cells coloured according to their preferred head direction). Landmark cells provide direct input to the corresponding head direction cells (upper right; shade of blue represents the firing rate). Left and right rotation cells (lower panels) make asymmetric recurrent connections with head direction cells, projecting strongly to either the left or right neighbour of the head direction cell from which they receive input. Upper right and lower images in panel **a** are adapted from ref. 48, Springer Nature Ltd. Panel **b** is adapted from ref. 50, CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Panel **c** is adapted with permission from refs. 46,51, AAAS and Annual Reviews.

rodents[46]. Thus, it is clear that a neural manifold does not uniquely determine the biophysical details of the circuit connectivity, highlighting the need for experimental measurements of connectivity to evaluate the candidate biophysical mechanisms suggested by theoretical models.

Complete circuit reconstruction

One way to precisely link a neural manifold to its underlying circuit is to directly measure the activity and anatomical connectivity of all of the neurons in the relevant circuits. In flies, such circuit dissection confirmed – with astonishing precision – the predictions of the ring attractor models described above[49,51,52,60]. Powerful experimental techniques, including RNA profiling and connectivity reconstruction through electron microscopy, enabled the identification of many cell types in the *D. melanogaster* central complex and a description of their anatomical connectivity with single-cell resolution[47,61]. This analysis revealed connectivity among many cell types in the fly head direction system, allowing the ring attractor model to be tested directly. It was found that local excitation between E-PG neurons sustains the activity bump in the ellipsoid body. The rotation cells predicted by the model were identified as P-EN neurons located in another neuropil called the protocerebral bridge. The recurrent connectivity pattern between EP-G and P-EN neurons was found to be asymmetric, meaning that P-EN neurons in the left and right sides of the protocerebral bridge rotate the position of the EP-G activity bump clockwise or anticlockwise, respectively[47,50]. These findings confirm precisely the angular velocity integration mechanism suggested in the ring attractor models. Many other cell types were also identified and their connectivity and function dissected[47], revealing a more complex anatomical structure than had been predicted by minimal circuit models. In turn, this led to the development of refined models incorporating the discovered anatomy[50,57,62–64].

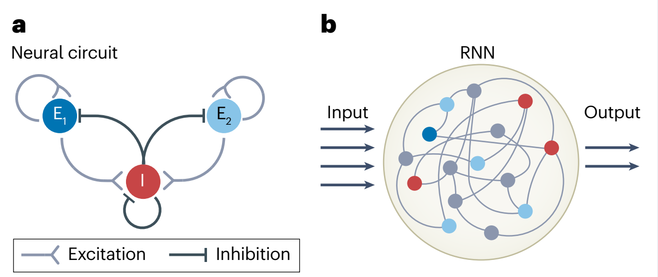
Box 2

Neural circuits

A neural circuit is a network of nodes in which the connections, dynamics of each node's activity and inputs are specified. According to the traditional view, the nodes in the network can be individual neurons, clusters of neurons or brain areas. The connections between nodes implement neural computations. For example, recurrent self-excitation in two clusters of excitatory neurons and cross-inhibition mediated by a third cluster of inhibitory neurons (see the figure, part **a**) can generate 'winner-take-all' dynamics, in which two discrete attractors support decision-making[10,18]. A network in which connectivity is spatially arranged in a pattern with local excitation and global inhibition can generate a continuous attractor (such as a ring attractor), in which a localized subset of active neurons ('bump') within the spatially organized network represents a continuous task variable (such as head direction, see Fig. 1b,c).

Traditional neural circuit models implement interpretable mechanisms that can relate connectivity to dynamics and behaviour. Interpretable mechanisms enable us to predict the behavioural effects of specific circuit perturbations, making it possible to experimentally test circuit mechanisms[21–25]. However, the simple connectivity structure of these traditional circuit models results in homogeneous tuning (for example, all neurons within a cluster have the same response profile), which is inconsistent with the distributed mixed selectivity (Box 1) that has been observed in brain recordings.

Artificial recurrent neural networks (RNNs) are a class of neural circuit models that can account for distributed mixed selectivity. An RNN consists of many recurrently connected units, with the weights of the connections optimized to produce a desired output from a specific external input (see the figure, part **b**). RNNs can be trained to reproduce behavioural responses in a cognitive task[111,112,116–119], low-dimensional manifolds[142] or recorded brain activity (in which case the activity of each RNN unit tracks a target experimental neuron[127]).



Distributed mixed selectivity emerges in RNNs through training[111]. However, trained RNN connectivity appears complex and the circuit mechanism that generates task-relevant dynamics in these networks is not immediately interpretable. Two RNNs trained to perform the same task may produce similar low-dimensional responses but have distinct high-dimensional connectivity[117]. Finding an interpretable connectivity structure that generates low-dimensional responses is important because it can allow us to determine whether different RNNs implement similar circuit mechanisms and to design perturbations to causally test these mechanisms.

Despite these challenges, there is evidence that interpretable circuit mechanisms can exist in networks with distributed mixed selectivity. The connectivity of such networks can contain a low-dimensional structure that implements casual interactions between distributed activity patterns on the manifold[120,121,124,139], similar to the interactions between nodes in a low-dimensional latent circuit (Box 4). This low-dimensional connectivity structure can be added to random connectivity in RNNs[89,120], making it challenging to identify. However, traditional circuit models can guide the search for interpretable mechanisms in RNNs[89,138].

The fruit fly head direction system is a unique example of a situation in which theoretical predictions made more than 20 years ago were confirmed experimentally at the level of single-cell connectivity, establishing a perfect correspondence between the circuit structure, dynamics on the neural manifold and behaviour. Although neural responses in the rodent head direction system organize on a similar ring manifold, they show more heterogeneity and lack a clear spatial topography. The precise circuit mechanism that generates dynamics on the ring manifold in the rodent head direction system is thus yet to be discovered.

Towards convergence in grid cells

Grid cells in the mammalian medial entorhinal cortex (MEC) provide an example of a navigational system for which the manifold and circuit perspectives have begun to converge, although their correspondence has not yet been established directly. MEC grid cells encode an animal's location in space and update this representation using information about the animal's speed and direction of motion, a computation known as path integration[65]. Spatial location in a flat environment is a two-dimensional variable that does not, per se, imply a periodic code. However, grid cells have been shown to represent space with a remarkably regular periodic pattern[66] (Fig. 2a). Grid cells activate whenever the animal's position coincides with any vertex of a regular grid of equilateral triangles spanning the environment. Grids of neighbouring cells share the same orientation and spacing but differ in their vertex locations (their phases). Across the MEC, grid cells cluster into a small number of anatomically overlapping modules with distinct scales and orientations of grids[67]. Within a grid module, phase relationships between pairs of grid cells are conserved in different environments despite extensive deformations of single-cell tuning[68]. Similar to the head direction system, the grid cell map is anchored to external landmarks but persists in their absence[69].

A toroidal manifold for spatial position

The grid-like periodicity of spatial tuning in single grid cells suggests that their population responses will organize on a low-dimensional manifold. As the tuning of single cells within a module varies only by spatial phase, it is expected that their population responses form a

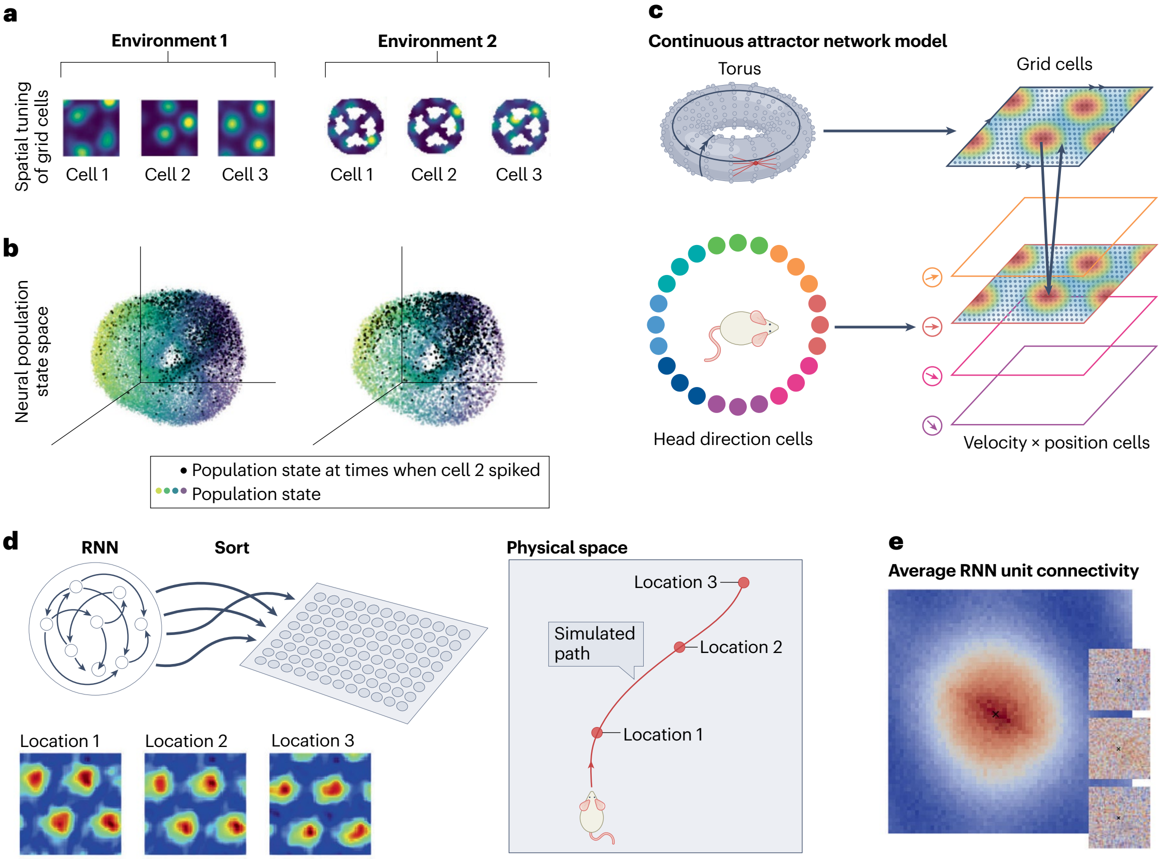


Fig. 2 | Manifolds and circuits for spatial position encoding. a, Responses of three example grid cells from a single grid module (in which cells have similar spatially periodic activity) in a freely moving rat in two different spatial environments. The firing rate maps indicate a reduced periodicity of spatial tuning in the second environment[71]. **b**, Despite changes in the spatial tuning of individual grid cells, neural population responses organize on the same toroidal manifold in different environments. Each point in the neural population state space represents the population activity state at a single time (dots coloured by first principal component of neural responses). Black dots indicate the population state at times when cell 2 fires. The clustering of black dots at the same location on the toroidal manifold indicates that there are stable relationships between the activity of grid cells across environments, which suggests that the manifold arises from a recurrent connectivity structure as in continuous attractor models[68]. **c**, A two-dimensional continuous attractor network model accounts for the toroidal manifold and path integration dynamics in grid cells[77,78]. The model consists of a network of grid cells spatially arranged on a two-dimensional torus (upper left; dots represent cells, red lines show connections made by one example cell). We can unfold the toroidal network into a two-dimensional sheet, in which the cells at opposite boundaries connect to each other (upper right; arrows mark connected boundaries). In this network, multiple focused areas (bumps) of activity form spontaneously, with their spatial pattern corresponding to co-active grid cells in the population. As the animal moves, these activity bumps move across the two-dimensional network to update the representation of spatial location via path integration. The movement of these bumps is mediated by local asymmetric connections between the grid cells and additional two-dimensional layers of cells that encode both head direction, speed and position (lower right). These layers of cells receive direct input from head direction cells (lower left) and are analogous to rotation cells in the ring attractor model (see Fig. 1). **d**, Grid-like responses can emerge in artificial recurrent neural networks (RNNs) trained to perform path integration[88,89]. Initially, RNN units are not arranged in any space; however, after training, they can be sorted on a two-dimensional sheet (upper left) so that units with similar phases of their grid tuning are close in space. As the RNN follows a simulated path through an environment (right panel), stable activity patterns on the two-dimensional neural sheet reveal multiple bumps with a topographic hexagonal grid activation (lower left, shown at three distinct locations along the simulated path), similar to that seen in classical attractor models. **e**, The connections from an individual RNN unit to its neighbours on the neural sheet appear unstructured (inset; connections made by three example units are shown, excitatory connections are red and inhibitory connections are blue). However, when averaged across many units, the connectivity reveals a structure in which there is local excitation (red) and global inhibition (blue), matching the mechanism for generating continuous attractor dynamics proposed in hand-crafted models. Parts **a** and **b** are adapted from ref. 71, Springer Nature Ltd. Part **c** is adapted from ref. 20, Springer Nature Ltd. Parts **d** and **e** are adapted with permission from ref. 89, Elsevier.

two-dimensional torus in state space[70]. Intuitively, the population response of grid cells within a module repeats cyclically whenever the animal moves along one of the two directions defining the grid period.

Evidence for such a torus-like manifold structure was discovered recently in a study that used Neuropixels probes to obtain simultaneous recordings from thousands of MEC grid cells in freely moving or sleeping rats[71] (Fig. 2b). This work showed that, as the animal moves in an open field, the population activity within a module also moves continuously across the toroidal manifold, updating the spatial representation via path integration[71]. The same toroidal manifold persists in the absence of sensory input and is maintained, with minimal distortion, across different environments and behavioural states from wakefulness to sleep[71]. Moreover, mapping the population activity at each time to a point on the identified torus in state space showed that individual cells fired preferentially when population responses fell at a particular location on this torus (Fig. 2b). Thus, individual grid cells are tuned to specific locations on the toroidal manifold and these locations do not change across environments[71] (Fig. 2b).

The crystalline rigidity of the toroidal manifold indicates that it is likely to arise from the connectivity structure in the circuit, and not from external input. Multiple tori (one in each module) observed experimentally[71] could arise from multiple subnetworks with toroidal topology[20]. In contrast to the head direction system, in which head direction is naturally encoded as a circular variable, this toroidal structure does not reflect the topology of the underlying spatial variable and, thus, its functional significance is still debated. One possibility is that the representation of position with respect to multiple tori with distinct periods provides a high-capacity combinatorial encoding that is read out downstream by place cells[72].

Discovering structure with manifold analyses. The example of grid cells shows that the tuning functions of single cells to external variables contain the same information as the manifold obtained from their trial-averaged responses[34]. However, manifold analyses can also reveal structure in neural population activity in situations in which estimating single-cell tuning is not possible or when the full set of variables encoded in neural activity are unknown. For example, during sleep, estimating single-cell tuning curves is not possible because there are no behavioural variables to which neural activity can be referenced. However, manifold analyses reveal the same manifolds in head direction cells[48,54] and grid cells[71] during sleep as during wakefulness, suggesting that they arise from anatomical connectivity. Similarly, manifold analyses enabled the discovery of a head direction circuit in the anterior hindbrain of larval zebrafish by demonstrating that neural responses in this area form a ring manifold[73]. As the fish were head fixed for volumetric calcium imaging, computing single-cell tuning to head direction was not possible. Moreover, unsupervised manifold discovery methods can reveal neural representations encoding an expansive task knowledge beyond external physical variables[74,75]. In the hippocampus of rodents performing decision-making tasks, neural population activity was well described by a low-dimensional manifold, within which both spatial location and abstract variables (such as accumulated evidence) were encoded in an orderly fashion. The manifold thus formed a conjoined cognitive map of the task[74], with some dimensions reflecting information beyond the measured behavioural variables[75]. Although the activity of single place cells in the hippocampus is known to vary substantially across repeated trials, manifold analysis has shown that this variability could result from neural trajectories taking different paths on the manifold and may reflect the operation of internal cognitive processes[75]. Thus, manifold analyses enable scientific insights beyond those made possible by the tuning-curve approach.

Circuit mechanisms for the toroidal manifold

Continuous attractor models. Similar to the head direction system, continuous attractor models provide a candidate circuit mechanism that can support the toroidal manifold and dynamics for path integration in grid cells[76]. These models are a direct extension of the one-dimensional ring attractor models into two dimensions. In a two-dimensional continuous attractor network model of grid cells, the cells are spatially arranged on a two-dimensional torus, with the strength of the recurrent excitatory connections between the cells decreasing in proportion to the distance that separates them[77,78] (Fig. 2c). In these networks, a single bump or multiple bumps of activity form spontaneously in a spatial pattern corresponding to the positions of co-active grid cells in the population, consistent with topographic organization of grid cells in the MEC[79–81]. This pattern of activity moves across the two-dimensional network to update the representation of spatial location via path integration according to the self-motion input[78]. The movements of activity bumps arise from local asymmetric connections between the grid cells and additional two-dimensional layers of cells that represent a combination of velocity and position, analogous to the layer of rotation cells in the head direction model[78] (Fig. 2c). Grid firing patterns can also arise in feedforward models in which spatial selectivity is inherited from external inputs[82–84], but this mechanism is inconsistent with the rigidity of the toroidal manifold across environments and behavioural states. By contrast, the spatially arranged connectivity in attractor models naturally leads to a rigid manifold structure that does not change with varying input[77,78,85].

Toroidal manifolds can arise in continuous attractor models that include either periodic boundary conditions (such that neurons at one boundary connect to neurons at the opposite boundary of the two-dimensional network, forming a torus) or aperiodic boundary conditions[76,85], but which applies to the grid cell system is unknown. Unlike the fly head direction system, the anatomical connectivity supporting the toroidal manifold in grid cell population activity remains unknown, with such research being hampered by a lack of simple topography and the more limited set of circuit dissection tools currently available in mammals. In the absence of direct connectivity measurements, the alternative circuit models could be evaluated by testing their predictions in experiments that combine sparse neural recordings with global perturbation strategies[76]. In particular, perturbations of either the time constant of neurons or the gain of the recurrent inhibition between neurons have predictable effects on the spatial tuning relationships between pairs of cells (and hence the manifold) in candidate models, and these effects can be detected using only a small number of neurons[76]. Thus, cortical cooling to alter the time constant[86] and drug infusions to alter the gain of recurrent inhibition[87] are two feasible experimental manipulations that could, in principle, be used to test alternative circuit models of grid cells in future studies.

Emergent circuit mechanisms in complex networks. To help us understand the links between the circuit structure and neural dynamics in mammalian navigational systems that have more complex and heterogeneous topography, we can turn to artificial recurrent neural network (RNN) models of path integration in which connectivity is not topographically arranged. RNNs can be trained to perform path integration by optimizing recurrent connectivity parameters. In these networks, representations can emerge that are similar to those formed by biological head direction cells and grid cells[59,88,89]. In the hand-crafted attractor models described above, the ring or toroidal response manifolds arise from connectivity that is spatially arranged in one or two dimensions, respectively. By contrast, RNN units are not arranged in any space, and ring and toroidal manifolds exist without topographic organization of neural responses. The trained RNN connectivity appears complex and the mechanism through which it generates precisely organized neural responses is not immediately obvious. Uncovering the mechanism for path integration in the RNN connectivity required analysis methods that arrange neurons in space according to their functional properties[59,89] (Fig. 2d). These analyses revealed a hidden structure in the RNN connectivity that matches the mechanism used for path integration in the hand-crafted attractor models. In both hand-crafted models and RNNs, the precisely organized spatial responses arise from a similar low-dimensional connectivity structure; however, in RNNs this structure is additively superimposed over random connectivity and its presence is therefore not obvious in the connectivity of individual units[59,89] (Fig. 2e). This means that, without the prior intuition provided by the theoretical models, finding this low-dimensional connectivity structure would have been extremely challenging. This example therefore demonstrates a more general point: although measuring the activity and anatomical connectivity of the same neurons can enable direct testing of hypothesized circuit mechanisms with single-cell resolution[89,90], without the guidance of theoretical models there is no universal path by which we may discover new mechanisms from the high-dimensional heterogeneous data. Therefore, theory is crucial for elucidating circuit mechanisms from simultaneous measurements of activity and connectivity.

Circuits with mixed selectivity

In contrast to navigational systems, the relationship between the neural manifold and circuit connectivity is more elusive in higher cortical areas that support cognitive functions, such as working memory[91–94] or decision-making[95,96]. Tasks used to study cognitive functions in animal experiments often have a simple topological structure (Box 1), akin to spatial navigation tasks. For example, a common visual spatial working memory task requires an animal to remember a location at a particular angle around a circle on a screen[97,98] (Fig. 3a). The remembered angle is a one-dimensional circular variable, just like the head direction. Similarly, in many decision-making tasks, task variables have a simple branching topology, in which diverging values of a decision variable represent alternative choices[95]. However, unlike neurons in navigational circuits, cortical neurons exhibit more complex and heterogeneous responses in these tasks, with a less obvious link to neural computation and circuit connectivity.

Classical circuit models of cognitive tasks

Early studies using single-cell recordings focused on the salient, interpretable tuning features of single neurons that aligned with the task structure. During working memory maintenance, for example, some neurons in the primate prefrontal cortex (PFC) show persistent activity with stimulus-dependent tuning, providing an essential neural correlate of working memory[91,93,94,97,98]. Similarly, during decision-making, the firing rates of single neurons across many cortical areas tend to ramp up or down, diverging across trials on which the animal makes different choices[95,99–104].

These lucid features of single-neuron responses map naturally to the activity in attractor network models with simple connectivity structures. A ring attractor model (with the same connectivity as the ring attractor model of the head direction cells) captures the stimulus-dependent persistent activity observed during spatial working memory tasks[9,105,106] (Fig. 3a). Moreover, the continuous attractor dynamics in this model predict the relationship between the precision of a memory report and fluctuations in PFC activity[107] as well as memory deficits arising from circuit disruptions (such as altered excitation–inhibition balance) in mental illness[21,108–110]. The ramping activity associated with decision-making arises in discrete attractor models[10,15–18]. These networks include several groups of excitatory neurons, one for each choice, with stronger recurrent excitation within a group than across groups. The inhibitory neurons in these networks mediate winner-take-all competition between the excitatory populations so that in response to a stimulus, one group elevates its firing rate representing the decision outcome. The discrete attractor models predict the changes in ramping activity and decision-making behaviour that occur when optogenetic perturbations of neural activity are performed in rodents[23–25].

Manifolds for cognitive tasks

Recent large-scale recordings have exposed the rich complexity and heterogeneity of single-neuron responses in higher cortical areas and revealed that simple interpretable tuning, such as that described above, is a rare exception[27–30,111,112]. Only a relatively small fraction (5–10%) of PFC neurons show strictly tonic persistent activity during working memory, with all other neurons displaying complex temporal variations[28,29] (Fig. 3b). The responses of PFC neurons are even more perplexing in tasks that involve interactions between multiple variables, such as context-dependent decision-making[111,112]. Single neurons show mixed selectivity, responding to combinations of multiple task variables, and the encoding of those variables is distributed across the entire neuronal population and varies over time[27,30]. This distributed mixed selectivity (Box 1) does not fit with the classical attractor models, in which neurons inherit homogeneous tuning properties from clustered or spatially organized connectivity structures. The question that therefore arises is how neural computations should be understood in networks with distributed mixed selectivity.

Manifold analysis approaches this question by finding low-dimensional representations of task variables in the population state space that are not obvious in the heterogeneous responses of single neurons. To identify the manifold structure in neural response data, many dimensionality reduction methods model heterogeneous responses as linear combinations of a few latent variables to extract a low-dimensional subspace within the population state space in which task-related dynamics can be observed[31,113–115]. The manifold structure found within this low-dimensional subspace often agrees with the topology of task variables. For example, in the spatial working memory task described above, the high-dimensional PFC population activity contains a low-dimensional subspace in which stimulus representations are stable across time and arranged on a circle representing the remembered location[29] (Fig. 3b). The variation of population activity over time occurs in an orthogonal subspace, and therefore does not interfere with the stable mnemonic representation. In decision-making tasks, low-dimensional projections of neural population activity reveal branching trajectories that diverge at each decision point[111,112] (Fig. 3c).

Manifold analyses have uncovered manifolds that mirror the topology of task variables in many cortical areas and cognitive tasks[27–30,111,112]. Manifolds and single-neuron heterogeneity qualitatively similar to those observed in brain recordings also emerge in RNNs trained to perform cognitive tasks by optimizing recurrent connectivity parameters[111,112,116–119] (Box 2). These discoveries led to the proposition that computation through dynamics on a manifold may be a general principle for the organization of heterogeneous neural responses in the brain[40]. Moreover, it has been suggested that understanding computation on the level of connectivity and circuits may be unnecessary or even intractable in systems with distributed mixed selectivity[41]. Is such a strong proposition warranted, or can the manifold and circuit perspectives be reconciled despite single-cell heterogeneity?

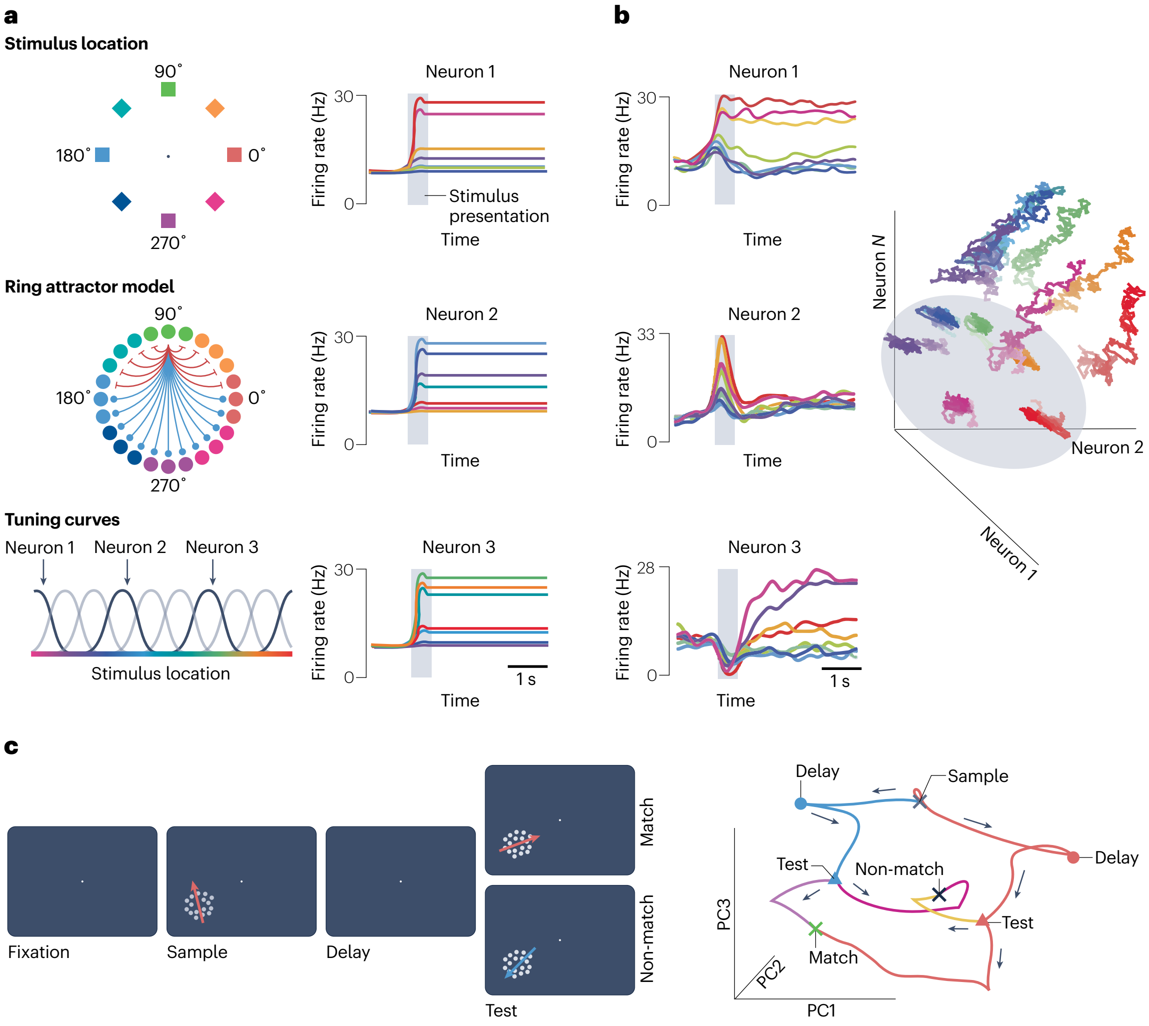


Fig. 3 | Low-dimensional task manifolds in heterogeneous responses of cortical neurons. a, A visual spatial working memory task requires an animal to remember the location of a stimulus positioned at a particular angle around a circle on a screen (upper left; spatial cues coloured according to their angle). A ring attractor model (middle left) in which there is strong local excitation (red lines) and global inhibition (blue lines) between model neurons results in stimulus-dependent persistent activity in the model neurons[9,105,106]. A representation of this activity in three model neurons in response to a stimulus in different locations (location indicated by colour) is shown in the right panels (*y* axis represents firing rate in Hertz), based on the findings reported in ref. 9. This model produces tuning curves for individual model neurons with identical shapes that uniformly cover the stimulus space (lower left). **b**, In the spatial working memory task, responses of single neurons recorded in the primate prefrontal cortex (PFC) show heterogeneous temporal profiles and stimulus tuning (left panels; responses of three example PFC neurons, using same colour coding as part **a**)[29]. The PFC population activity contains a linear subspace (right; grey circle) with stable stimulus encoding on a ring, where the position on the ring represents the remembered location. The population activity varies over time in an orthogonal non-coding subspace (right; coloured trajectories) without interfering with this stable mnemonic representation. The responses of single PFC neurons are heterogeneous because the coding and non-coding subspaces are rotated with respect to neural axes in the population state space. **c**, A delayed match-to-category task requires an animal to indicate whether a test stimulus belongs to the same category as a previously shown sample stimulus (left panel). This task involves two sequential decisions: what is the sample category and does it match the test category? In recurrent neural networks (RNNs) trained to perform this task, population responses form a branching manifold, which is visualized by projecting responses of RNN units onto the first three principal components[112] (right panel). Colours indicate task conditions corresponding to different pairs of sample and test categories. During the sample period, trajectories diverge (red and blue circles) and then approach one of two distinct states representing the memory of the sample category during the delay (red and blue triangles). During the test period, trajectories again diverge towards two other states representing match or non-match decisions (green and black crosses). Similar branching manifolds are observed in the population activity of neurons recorded in the lateral intraparietal area and PFC as animals complete the task[112]. Part **b** is adapted with permission from ref. 29, PNAS. Part **c** is adapted with permission from ref. 112, Elsevier.

Linking manifolds to connectivity

Low-rank recurrent neural networks. It is possible to reconcile the manifold and circuit perspectives by engineering nonlinear high-dimensional RNNs in which dynamics on a low-dimensional manifold arise from low-dimensional connectivity (Box 3). In these RNNs, connectivity is constructed to be low rank[120,121]. Rank-one connectivity is an outer product of two high-dimensional vectors *m* and *n* of length *N* (where *N* is the number of neurons in the high-dimensional network).

Box 3

Dimensionality of manifolds and circuits

Dimensionality generally refers to the number of independent variables that are necessary to describe an object, such as a neural manifold. We can define different types of dimensionality depending on the choice of these variables. The linear dimensionality (sometimes called embedding dimensionality[42]) is the smallest number of orthogonal directions that span a linear subspace containing the manifold. The nonlinear intrinsic dimensionality is the minimal number of continuous variables necessary to parameterize the manifold. For example, the intrinsic dimensionality of the ring manifold is one, because it can be parameterized by a single angular variable (Box 1). The linear dimensionality of the ring manifold in the neural state space depends on the width of the tuning curves of individual neurons and can be very high if the tuning curves are narrow[34].

The dimensionality of a neural manifold depends, in part, on the connectivity of the underlying neural circuit. Such connectivity can be described by a matrix J in which the elements J_{ij} specify the weight of connection from neuron j to neuron i. A connectivity matrix can also contain a low-dimensional interpretable structure. One example of low-dimensional connectivity is low-rank connectivity. The rank of a connectivity matrix is the number of orthogonal vectors needed to reconstruct the matrix. A connectivity matrix is low rank if its N columns (or rows) can be assembled as linear combinations of a much smaller number $k \ll N$ of columns (or rows)[143]. The simplest possible type of low-rank connectivity is a rank-one connectivity matrix J, which is fully specified by two N-dimensional vectors, $m = \{m_i\}$ and $n = \{n_j\}$ (i and j are indices taking values from 1 to N; see the figure, part **a**). Every column of a rank-one matrix J is a multiple of the vector m, and every row is a multiple of the vector n (that is, J is an outer product of m and n):

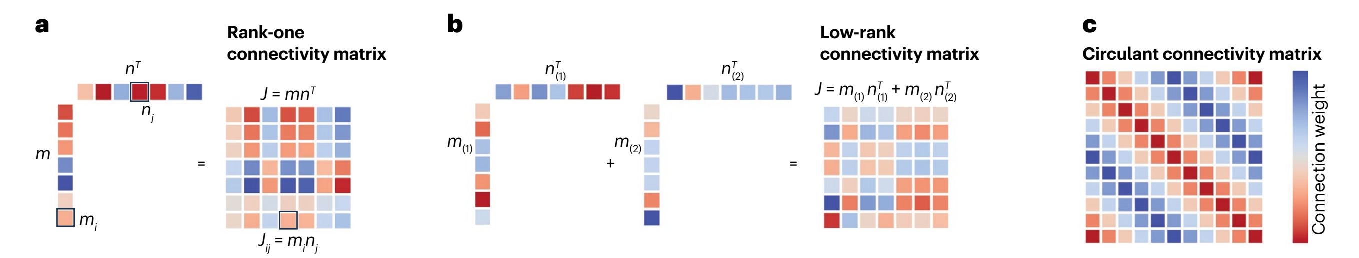
$$J = mn^T, \quad J_{ij} = m_i n_j.$$

One way to compose a low-rank connectivity matrix is by adding together k rank-one terms (see the figure, part **b**):

$$J = \sum_{l=1}^{k} m_{(l)} n_{(l)}^T.$$

Another example of a low-dimensional interpretable connectivity structure is the circulant connectivity matrix that is used in a ring attractor model. The circulant matrix is fully specified by a single vector n that defines the connectivity profile of one neuron[143]. Each row of a circulant matrix is obtained through a circular shift of the vector n one element to the right relative to the preceding row (see the figure, part **c**). Although this matrix can be high rank if the connectivity profile is narrow, it contains an interpretable low-dimensional structure.

It is unknown whether low-dimensional manifolds always arise from low-dimensional connectivity, or whether they can emerge without any low-dimensional structure in either the input or recurrent connectivity. Many dimensionality reduction methods exist for finding low-dimensional manifolds in neural population activity, but there are no general approaches for finding the corresponding low-dimensional structure in the connectivity. In this Perspective, we highlight examples in which it has been possible to establish a relationship between the low-dimensional activity and connectivity[55,76,89,120,124]. Finding such cases is important because they reveal interpretable circuit mechanisms that can be validated in perturbation experiments.


a
Rank-one connectivity matrix
n^T
n_j
m
m_i
=
$J = mn^T$
$J_{ij} = m_i n_j$
b
$n^T_{(1)}$
$n^T_{(2)}$
$m_{(1)}$
+
$m_{(2)}$
=
Low-rank connectivity matrix
$J = m_{(1)}n^T_{(1)} + m_{(2)}n^T_{(2)}$
c
Circulant connectivity matrix
Connection weight

With this connectivity structure, network dynamics are confined to the two-dimensional subspace of the neural activity space spanned by vectors m and n when external input is aligned with n. Moreover, activity along direction n drives activity along direction m, creating a substrate for implementing computations. By composing low-rank connectivity from several rank-one terms (Box 3), it is possible to construct RNNs with dynamics flowing on manifolds spanning a few directions in neural activity space, designed to solve various cognitive tasks[120,122]. It is, however, important to note that, although low-rank RNNs can be engineered to solve certain tasks, it does not necessarily follow that brain networks use this particular connectivity structure to solve the same tasks. Low-rank solutions are one of many candidate mechanisms, and it remains unknown whether low-dimensional connectivity underlies manifolds in other heterogeneous networks, such as task-optimized RNNs or the brain. Moreover, it is unclear whether low-rank RNNs utilize mechanisms similar to classical circuit models or implement truly novel solutions that are emergent in high-dimensional nonlinear systems.

Latent circuits in heterogeneous networks. We can start to gain a better understanding of the link between low-rank connectivity in large recurrent networks and low-dimensional circuits with few populations interacting via excitation and inhibition through an examination of low-rank linear dynamical systems[123]. Linear dynamical systems provide a mathematically tractable approximation to nonlinear brain dynamics[111]. In a high-dimensional linear network with low-rank connectivity composed of a few orthogonal vectors $q_{(i)}$, the dynamics are confined to a low-dimensional subspace spanned by these vectors[123] (Box 4). Each vector $q_{(i)}$ specifies a direction in the high-dimensional state space. These vectors, assembled as columns into the orthonormal matrix Q, define the low-dimensional subspace in which the dynamics unfold. Mathematically, these dynamics correspond to a high-dimensional embedding of a low-dimensional dynamical system, in which recurrent connectivity captures causal interactions between directions $q_{(i)}$ along the manifold. Thus, the low-dimensional dynamical system is latent in the high-dimensional network[31], and we call it the latent circuit[122,124]. The embedding matrix Q provides a mapping between the latent circuit and the high-dimensional network such that a latent node i maps onto a direction $q_{(i)}$. A connection between two latent nodes maps onto a distributed connectivity pattern given by an outer product of two corresponding directions along the manifold. This mapping allows us to translate perturbations of activity and connectivity from the low-dimensional circuit onto the high-dimensional network, making it possible to causally test the circuit mechanism. Thus, in linear networks, we can formally link the manifold and circuit perspectives with a latent circuit structure that constrains dynamics on the manifold[122,124].

To what extent does the intuition gained from such linear systems extend to nonlinear recurrent networks such as RNNs or the brain? A recent preprint[124] developed an approach for fitting high-dimensional neural responses with a model in which these responses arise as an embedding of low-dimensional dynamics generated by a nonlinear latent circuit. Using this approach, the latent circuit connectivity and the embedding matrix can be simultaneously inferred from neural responses, making it equally applicable to the activity generated by an RNN or the brain. In general, however, it is unclear whether a low-dimensional circuit can satisfactorily account for the responses of a high-dimensional network. If the solutions to cognitive tasks that emerge in high-dimensional systems are qualitatively different from those that operate in small circuits, then a low-dimensional circuit

Box 4

Latent circuits in linear low-rank networks

Linear dynamical systems provide a simple case in which we can link low-rank networks with the classical picture of a few nodes interacting via excitatory and inhibitory connections in a circuit. Consider a low-dimensional linear dynamical system:

$$\dot{x} = Ax, \tag{1}$$

which we can view as a circuit with a small number n of nodes interacting via an $n \times n$ recurrent connectivity matrix $A = \{A_{ij}\}$ (see the figure, part **a**). We can embed these low-dimensional dynamics in a high-dimensional state space to construct a high-dimensional linear dynamical system in which this circuit is latent. We can do this by considering the N-dimensional variable $y = Qx$, where Q is an orthonormal embedding matrix of shape $N \times n$ ($n \ll N$)[143] (see the figure, part **b**). The dynamics of y are then described by a linear dynamical system:

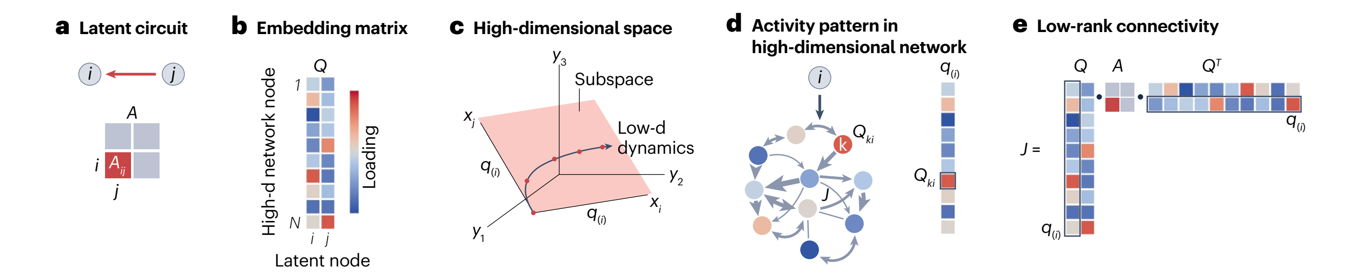
$$\dot{y} = Jy, \tag{2}$$

with the low-rank connectivity matrix $J = QAQ^T$ of size $N \times N$[144]. The dynamics of this high-dimensional network are confined to the low-dimensional subspace spanned by the columns of Q and follow Eq. (1) within this subspace (see the figure, part **c**). From this perspective, the low-dimensional dynamical system is latent in the high-dimensional network y, with dynamics described by latent variables x. The activity of the node x_i in the low-dimensional circuit maps to a high-dimensional activity pattern aligned with the vector $q_{(i)}$, where $q_{(i)}$ is the ith column of Q (see the figure, part **d**). Moreover, writing J as a sum of rank-one terms (Box 2):

$$J = \sum_{ij} A_{ij} q_{(i)} q_{(j)}^T, \tag{3}$$

we see that an edge from node j to node i in the latent circuit maps to the outer product $q_{(i)} q_{(j)}^T$ in the connectivity of the high-dimensional network (see the figure, part **e**). In this way, node j driving node i in the latent circuit corresponds to the activity along $q_{(j)}$ driving the activity along $q_{(i)}$ via low-rank connectivity[124].

Recent theoretical work suggests that low-dimensional dynamics can arise from low-rank connectivity in nonlinear RNNs[120–122,124,139]. Under what conditions low-dimensional manifolds arise from low-dimensional connectivity in nonlinear recurrent networks is an open question. Finding such cases will enable validation of distributed circuit mechanisms via perturbations[124].



model should not be able to adequately account for the task-related dynamics of the large system. However, if the dynamics of the large system are accurately predicted by the low-dimensional circuit model, then this would suggest that the low-dimensional circuit mechanism may be latent in the high-dimensional system. Applied to RNNs optimized to perform a cognitive task, this approach revealed a low-dimensional circuit structure in these networks, which was validated by patterned perturbations of the RNN activity and connectivity[124]. Future work is needed to determine under what conditions such low-dimensional connectivity underlies the manifold structure in heterogeneous neural responses.

This latent circuit approach is similar to methods which fit neural responses with latent linear dynamical systems[125,126] but incorporates a biologically plausible nonlinearity and task-relevant inputs and outputs, and therefore provides an interpretable model of the dynamics supporting cognitive task execution. Other approaches estimate full high-dimensional RNN connectivity by fitting RNN models to neural response data[127–129]. However, the full high-dimensional connectivity is not uniquely constrained by low-dimensional neural trajectories and is not easily interpretable[89] (Box 2). By contrast, the latent circuit approach infers only the low-dimensional connectivity structure generating task-related dynamics and can be sufficiently constrained by low-dimensional neural responses. The latent circuit framework also provides a conceptual advance over other methods for fitting neural circuit models to neural response data[130–132] by introducing the idea that circuit mechanisms may be distributed across heterogeneous neural populations.

Demixed representations versus circuit mechanisms. Identifying task manifolds from neural response data requires dimensionality reduction to project high-dimensional activity onto a low-dimensional subspace. The resulting low-dimensional representations are not unique and depend on the choice of a dimensionality reduction method. Intuition tells us that there are an infinite number of ways to project high-dimensional data onto a low-dimensional subspace, and each projection yields a distinct view of the data manifold. For example, principal component analysis looks for the projections that account for the most variance in the response data, irrespective of whether this variance is related to the task execution. By contrast, targeted dimensionality reduction methods aim to identify task-related dimensions in neural population activity, usually by searching for pure (demixed) representations of task variables[30,111,113,114]. Demixing approaches find directions in the neural population state space that correlate with each task variable, so that representations of task variables do not interact and are demixed in orthogonal dimensions[30,111,113,114]. The objective of demixing task variables contrasts with the mechanistic perspective provided by neural circuit models, in which nodes representing task variables interact via recurrent connectivity. These interactions are crucial as they implement the computations necessary to solve the task. Thus, demixing approaches and dimensionality reduction incorporating mechanistic constraints can infer different task manifolds from the same neural responses. Indeed, dimensionality reduction based on regression and the latent circuit model approach yielded contradictory conclusions about the representations of irrelevant stimuli in RNN models of context-dependent decision-making, but only the representations identified by the latent circuit model were validated using causal perturbations[124]. These results thus highlight the importance of interpreting neural representations within the context of circuit mechanisms.

Outlook for mixed selectivity circuits

The recent theoretical work described above suggests that the organization of heterogeneous neural responses on a manifold arises from a low-dimensional connectivity structure in the circuit[89,120,122,124]. These theoretical insights open new possibilities for testing the circuit–manifold correspondence in future experiments. In particular, the use of latent circuit inference from neural activity data can generate specific mechanistic hypotheses about how neurons interact to produce behaviour. Experiments can test these causal relationships by validating the behavioural effects of patterned perturbations of neural activity and connectivity predicted by the latent circuit model. Patterned activity perturbations in behaving animals are becoming increasingly more feasible with advances in optogenetic stimulation[133–135]. Although patterned connectivity perturbations are still experimentally out of reach, neural recordings followed by detailed connectivity reconstructions can validate the circuit–manifold relationship predicted by the theory. It is likely that the anatomical structure will be more complex than predicted by the minimal circuit models and that distinct cell types may play specific functional roles, as in the *Drosophila* head direction system. These anatomical discoveries will lead to refinements of theoretical models to solidify the relationship between neural manifolds and circuits.

Conclusions and perspectives

Our understanding of how cognitive computations emerge from collective dynamics in neural populations has advanced significantly over the past decade. Neural manifolds were in the vanguard of many breakthroughs that have led to a conceptual shift in focus from the single neuron to the neural population[136,137]. Neural manifolds gracefully compress the daunting complexity and heterogeneity of single-neuron responses to reveal interpretable low-dimensional structure on the population level that can often be related to the computational scaffold of the behavioural task. The successes in describing neural computations as dynamics on low-dimensional manifolds spurred the idea that manifolds are the necessary and sufficient building blocks to explain cognition[40]. At its extreme, this view suggests that understanding cognitive functions on the level of connectivity and circuits is unnecessary and may even be impossible[41].

The work we reviewed here supports an alternative view that the manifold and circuit approaches to cognition are inseparable. Representations of task variables on low-dimensional neural manifolds mirror the topological structure of the cognitive task variables and, in many cases, manifolds arise from the low-dimensional connectivity structure in the circuit[49,68,89,120,122,124]. In navigational systems, the regularity of single-neuron responses and simple manifold geometry naturally suggest the underlying connectivity structure and form the basis for theoretical circuit models[50,51]. Measurements of the activity and connectivity of all neurons in the entire circuit provide a direct test of such models and fully confirm the ring attractor model in the fly navigational system[47]. The intuition provided by theoretical models is crucial for identifying circuit mechanisms from simultaneous measurements of activity and connectivity[89,90,138]. In the mammalian neocortex, the diversity of single-neuron responses conceals the link to the underlying connectivity. However, the organization of neural responses on a low-dimensional manifold can be related to a low-dimensional connectivity structure that gives rise to the manifold in neural activity space[120,122,124]. This relationship was directly confirmed in RNNs[124], and future work can test it in biological data.

Similar to the classical circuit models, the neural manifold and the circuit mechanism intertwine in systems with distributed mixed selectivity. However, a circuit mechanism in such systems exists not as specific connections between pairs of neurons but as a low-dimensional connectivity structure that enables one distributed activity pattern to influence another[120–122,124,139]. The activity patterns define the low-dimensional manifold, and the low-dimensional connectivity structure implements causal interactions between different dimensions on this manifold.

Why should we seek circuit structure and not satisfy ourselves with descriptions of neural computations as dynamics on manifolds? Without links to the underlying circuit mechanism, neural manifolds provide merely an abstract statistical description of the population dynamics. Thus, the dynamics discovered from data are sensitive to nuances of the statistical method. Different methods sometimes arrive at seemingly contradictory hypotheses without a clear path to falsify them. By contrast, a circuit mechanism captures causal interactions between neurons and generates testable predictions for perturbation experiments, offering an objective way to identify an accurate model among plausible circuits. Mechanistic understanding of the neural circuits underlying cognition will give us new opportunities to interface with these circuits and treat mental disorders. Therefore, neural manifolds are not the end goal but a link between experiments and theoretical modelling that is needed to identify the circuit mechanisms giving rise to the neural dynamics that drive behaviour.

Published online: 13 April 2023

References

1. Steinmetz, N. A., Koch, C., Harris, K. D. & Carandini, M. Challenges and opportunities for large-scale electrophysiology with Neuropixels probes. *Curr. Opin. Neurobiol.* **50**, 92–100 (2018).
2. Steinmetz, N. A. et al. Neuropixels 2.0: a miniaturized high-density probe for stable, long-term brain recordings. *Science* **372**, eabf4588 (2021).
3. Ebrahimi, S. et al. Emergent reliability in sensory cortical coding and inter-area communication. *Nature* **605**, 713–721 (2022).
4. Oh, S. W. et al. A mesoscale connectome of the mouse brain. *Nature* **508**, 207–214 (2014).
5. Markov, N. T. et al. Cortical high-density counterstream architectures. *Science* **342**, 1238406 (2013).
6. Harris, J. A. et al. Hierarchical organization of cortical and thalamic connectivity. *Nature* **508**, 207–230 (2019).
7. Huang, L. et al. BRICseq bridges brain-wide interregional connectivity to neural activity and gene expression in single animals. *Cell* **182**, 177–188.e27 (2020).
8. Skaggs, W. E., Knierim, J. J., Kudrimoti, H. S. & McNaughton, B. L. A model of the neural basis of the rat's sense of direction. *Adv. Neural Inf. Process. Syst.* **7**, 173–180 (1995).
9. Compte, A., Brunel, N., Goldman-Rakic, P. S. & Wang, X.-J. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb. Cortex* **10**, 910–923 (2000).
10. Wang, X. J. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* **36**, 955–968 (2002).
11. Machens, C. K., Romo, R. & Brody, C. D. Flexible control of mutual inhibition: a neural model of two-interval discrimination. *Science* **307**, 1121–1124 (2005).
12. Lo, C.-C. & Wang, X.-J. Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nat. Neurosci.* **9**, 956–963 (2006).
13. Engel, T. A. & Wang, X. J. Same or different? A neural circuit mechanism of similarity-based pattern match decision making. *J. Neurosci.* **31**, 6982–6996 (2011).
14. Ardid, S. & Wang, X.-J. A tweaking principle for executive control: neuronal circuit mechanism for rule-based task switching and conflict resolution. *J. Neurosci.* **33**, 19504–19517 (2013).
15. Roach, J. P., Churchland, A. K. & Engel, T. A. Choice selective inhibition drives stability and competition in decision circuits. *Nat. Commun.* **14**, 147 (2023).
16. Martí, D., Deco, G., Mattia, M., Gigante, G. & Giudice, P. D. A fluctuation-driven mechanism for slow decision processes in reverberant networks. *PLoS ONE* **3**, e2534 (2008).
17. Ksander, J., Katz, D. B. & Miller, P. A model of naturalistic decision making in preference tests. *PLoS Comput. Biol.* **17**, e1009012 (2021).
18. Wong, K.-F. & Wang, X.-J. A recurrent network mechanism of time integration in perceptual decisions. *J. Neurosci.* **26**, 1314–1328 (2006).
19. Wang, X.-J. in *Principles of Frontal Lobe Function* (eds Stuss, D. T. & Knight, R. T.) 226–248 (Oxford Academic, 2013).
20. McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I. & Moser, M.-B. Path integration and the neural basis of the 'cognitive map'. *Nat. Rev. Neurosci.* **7**, 663–678 (2006).
21. Murray, J. D. et al. Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model. *Cereb. Cortex* **24**, 859–872 (2014).
22. Lam, N. H. et al. Effects of altered excitation–inhibition balance on decision making in a cortical circuit model. *J. Neurosci.* **42**, 1035–1053 (2021).
23. Inagaki, H. K., Fontolan, L., Romani, S. & Svoboda, K. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature* **566**, 212–217 (2019).
24. Finkelstein, A. et al. Attractor dynamics gate cortical information flow during decision-making. *Nat. Neurosci.* **24**, 843–850 (2021).
25. Duan, C. A. et al. Collicular circuits for flexible sensorimotor routing. *Nat. Neurosci.* **24**, 1110–1120 (2021).
26. Urai, A. E., Doiron, B., Leifer, A. M. & Churchland, A. K. Large-scale neural recordings call for new insights to link brain and behavior. *Nat. Neurosci.* **25**, 11–19 (2022).
27. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).
28. Cavanagh, S. E., Towers, J. P., Wallis, J. D., Hunt, L. T. & Kennerley, S. W. Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. *Nat. Commun.* **9**, 3498 (2018).
29. Murray, J. D. et al. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc. Natl Acad. Sci. USA* **114**, 394–399 (2017).
30. Machens, C. K., Romo, R. & Brody, C. D. Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. *J. Neurosci.* **30**, 350–360 (2010).
31. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* **17**, 1500–1509 (2014).
32. Gallego, J. A., Perich, M. G., Miller, L. E. & Solla, S. A. Neural manifolds for the control of movement. *Neuron* **94**, 978–984 (2017).
33. Duncker, L. & Sahani, M. Dynamics on the manifold: identifying computational dynamical activity from neural population recordings. *Curr. Opin. Neurobiol.* **70**, 163–170 (2021).
34. Kriegeskorte, N. & Wei, X.-X. Neural tuning and representational geometry. *Nat. Rev. Neurosci.* **22**, 703–718 (2021).
35. Pandarinath, C. et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat. Methods* **15**, 805–815 (2018).
36. Duncker, L., Bohner, G., Boussard, J. & Sahani, M. Learning interpretable continuous-time models of latent stochastic dynamical systems. *Proc. 36th Intl Conf. Machine Learning* **97**, 1726–1734 (2019).
37. Genkin, M. & Engel, T. A. Moving beyond generalization to accurate interpretation of flexible models. *Nat. Mach. Intell.* **2**, 674–683 (2020).
38. Zhao, Y. & Park, I. M. Variational online learning of neural dynamics. *Front. Comput. Neurosci.* **14**, 71 (2020).
39. Genkin, M., Hughes, O. & Engel, T. A. Learning non-stationary Langevin dynamics from stochastic observations of latent trajectories. *Nat. Commun.* **12**, 5986 (2021).
40. Vyas, S., Golub, M. D., Sussillo, D. & Shenoy, K. V. Computation through neural population dynamics. *Annu. Rev. Neurosci.* **43**, 249–275 (2020).
41. Barack, D. L. & Krakauer, J. W. Two views on the cognitive brain. *Nat. Rev. Neurosci.* **22**, 359–371 (2021).
42. Jazayeri, M. & Ostojic, S. Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Curr. Opin. Neurobiol.* **70**, 113–120 (2021).
43. Knierim, J. J. & Zhang, K. Attractor dynamics of spatially correlated neural activity in the limbic system. *Annu. Rev. Neurosci.* **35**, 267–285 (2012).
44. Khona, M. & Fiete, I. R. Attractor and integrator networks in the brain. *Nat. Rev. Neurosci.* **23**, 744–766 (2022).
45. Clark, B. J. & Taube, J. S. Vestibular and attractor network basis of the head direction cell signal in subcortical circuits. *Front. Neural Circuits* **6**, 7 (2012).
46. Hulse, B. K. & Jayaraman, V. Mechanisms underlying the neural computation of head direction. *Annu. Rev. Neurosci.* **43**, 1–24 (2016).
47. Turner-Evans, D. B. et al. The neuroanatomical ultrastructure and function of a biological ring attractor. *Neuron* **108**, 145–163 (2020).
48. 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. *Nat. Neurosci.* **22**, 1512–1520 (2019).
49. Seelig, J. D. & Jayaraman, V. Neural dynamics for landmark orientation and angular path integration. *Nature* **521**, 186–191 (2015).
50. Turner-Evans, D. et al. Angular velocity integration in a fly heading circuit. *eLife* **6**, e23496 (2017).
51. Kim, S. S., Rouault, H., Druckmann, S. & Jayaraman, V. Ring attractor dynamics in the *Drosophila* central brain. *Science* **356**, 849–853 (2017).
52. Kim, S. S., Hermundstad, A. M., Romani, S., Abbott, L. F. & Jayaraman, V. Generation of stable heading representations in diverse visual scenes. *Nature* **576**, 126–131 (2019).
53. Ajabi, Z., Keinath, A. T., Wei, X.-X. & Brandon, M. P. Population dynamics of head-direction neurons during drift and reorientation. *Nature* https://doi.org/10.1038/s41586-023-05813-2 (2023).
54. Duszkiewicz, A. J. et al. Reciprocal feature encoding by cortical excitatory and inhibitory neurons. Preprint at *bioRxiv* https://doi.org/10.1101/2022.03.14.484357 (2022).
55. Zhang, K. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *J. Neurosci.* **16**, 2112–2126 (1996).
56. Redish, A. D., Elga, A. N. & Touretzky, D. S. A coupled attractor model of the rodent head direction system. *Netw. Comput. Neural Syst.* **7**, 671–685 (1996).

57. Cope, A. J., Sabo, C., Vasilaki, E., Barron, A. B. & Marshall, J. A. R. A computational model of the integration of landmarks and motion in the insect central complex. *PLoS ONE* **12**, e0172325 (2017).
58. Song, P. & Wang, X.-J. Angular path integration by moving "hill of activity": a spiking neuron model without recurrent excitation of the head-direction system. *J. Neurosci.* **25**, 1002–1014 (2005).
59. Cueva, C. J., Wang, P. Y., Chin, M. & Wei, X.-X. Emergence of functional and structural properties of the head direction system by optimization of recurrent neural networks. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1912.10189 (2019).
60. Green, J. et al. A neural circuit architecture for angular integration in *Drosophila*. *Nature* **546**, 101–106 (2017).
61. Wolff, T., Iyer, N. A. & Rubin, G. M. Neuroarchitecture and neuroanatomy of the *Drosophila* central complex: a GAL4-based dissection of protocerebral bridge neurons and circuits. *J. Comp. Neurol.* **523**, 997–1037 (2015).
62. Kutschireiter, A., Basnak, M. A., Wilson, R. I. & Drugowitsch, J. Bayesian inference in ring attractor networks. *Proc. Natl Acad. Sci. USA* **120**, e2210622120 (2023).
63. Lyu, C., Abbott, L. F. & Maimon, G. Building an allocentric travelling direction signal via vector computation. *Nature* **601**, 92–97 (2021).
64. Su, T.-S., Lee, W.-J., Huang, Y.-C., Wang, C.-T. & Lo, C.-C. Coupled symmetric and asymmetric circuits underlying spatial orientation in fruit flies. *Nat. Commun.* **8**, 139 (2017).
65. Mittelstaedt, M. L. & Mittelstaedt, H. Homing by path integration in a mammal. *Naturwissenschaften* **67**, 566–567 (1980).
66. Fyhn, M., Molden, S., Witter, M. P., Moser, E. I. & Moser, M.-B. Spatial representation in the entorhinal cortex. *Science* **305**, 1258–1264 (2004).
67. Stensola, H. et al. The entorhinal grid map is discretized. *Nature* **492**, 72–78 (2012).
68. Yoon, K. et al. Specific evidence of low-dimensional continuous attractor dynamics in grid cells. *Nat. Neurosci.* **16**, 1077–1084 (2013).
69. Hafting, T., Fyhn, M., Molden, S., Moser, M.-B. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**, 801–806 (2005).
70. Gao, R., Xie, J., Wei, X.-X., Zhu, S.-C. & Wu, Y. N. On path integration of grid cells: group representation and isotropic scaling. *Advances in Neural Information Systems 34* https://proceedings.neurips.cc/paper/2021/hash/f076073b2082f8741a9cd07b789c77a0-Abstract.html (2021).
71. Gardner, R. J. et al. Toroidal topology of population activity in grid cells. *Nature* **602**, 123–128 (2022).
72. Fiete, I. R., Burak, Y. & Brookings, T. What grid cells convey about rat location. *J. Neurosci.* **28**, 6858–6871 (2008).
73. Petrucco, L. et al. Neural dynamics and architecture of the heading direction circuit in a vertebrate brain. Preprint at *bioRxiv* https://doi.org/10.1101/2022.04.27.489672 (2022).
74. Nieh, E. H. et al. Geometry of abstract learned knowledge in the hippocampus. *Nature* **595**, 80–84 (2021).
75. Low, R. J., Lewallen, S., Aronov, D., Nevers, R. & Tank, D. W. Probing variability in a cognitive map using manifold inference from neural dynamics. Preprint at *bioRxiv* https://doi.org/10.1101/418939 (2018).
76. Widloski, J., Marder, M. P. & Fiete, I. R. Inferring circuit mechanisms from sparse neural recording and global perturbation in grid cells. *eLife* **7**, e33503 (2018).
77. Samsonovich, A. & McNaughton, B. L. Path integration and cognitive mapping in a continuous attractor neural network model. *J. Neurosci.* **17**, 5900–5920 (1997).
78. Conklin, J. & Eliasmith, C. A controlled attractor network model of path integration in the rat. *J. Comput. Neurosci.* **18**, 183–203 (2005).
79. Gu, Y. et al. A map-like micro-organization of grid cells in the medial entorhinal cortex. *Cell* **175**, 737–750.e30 (2018).
80. Heys, J. G., Rangarajan, K. V. & Dombeck, D. A. The functional micro-organization of grid cells revealed by cellular-resolution imaging. *Neuron* **84**, 1079–1090 (2014).
81. Obenhaus, H. A., Zong, W., Jacobsen, R. I. & Moser, E. I. Functional network topography of the medial entorhinal cortex. *Proc. Natl Acad. Sci. USA* **119**, e2121655119 (2022).
82. Kropff, E. & Treves, A. The emergence of grid cells: intelligent design or just adaptation? *Hippocampus* **18**, 1256–1269 (2008).
83. Dordek, Y., Soudry, D., Meir, R. & Derdikman, D. Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. *eLife* **5**, e10094 (2016).
84. Monsalve-Mercado, M. M. & Leibold, C. Hippocampal spike-timing correlations lead to hexagonal grid fields. *Phys. Rev. Lett.* **119**, 038101 (2017).
85. Burak, Y. & Fiete, I. R. Accurate path integration in continuous attractor network models of grid cells. *PLoS Comput. Biol.* **5**, e1000291 (2009).
86. Banerjee, A., Egger, R. & Long, M. A. Using focal cooling to link neural dynamics and behavior. *Neuron* **109**, 2508–2518 (2021).
87. Rudolph, U. & Möhler, H. Analysis of GABAA receptor function and dissection of the pharmacology of benzodiazepines and general anesthetics through mouse genetics. *Annu. Rev. Pharmacol. Toxicol.* **44**, 475–498 (2004).
88. Cueva, C. J. & Wei, X.-X. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1803.07770 (2018).
89. Sorscher, B., Mel, G. C., Ocko, S. A., Giocomo, L. M. & Ganguli, S. A unified theory for the computational and mechanistic origins of grid cells. *Neuron* **111**, 121–137.e13 (2022).
90. Bock, D. D. et al. Network anatomy and in vivo physiology of visual cortical neurons. *Nature* **471**, 177–182 (2011).
91. Wang, X.-J. 50 years of mnemonic persistent activity: quo vadis? *Trends Neurosci.* **44**, 888–902 (2021).
92. Fuster, J. M. & Alexander, G. E. Neuron activity related to short-term memory. *Science* **173**, 652–654 (1971).
93. Romo, R., Brody, C. D., Hernández, A. & Lemus, L. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* **399**, 470–473 (1999).
94. Kamiński, J. & Rutishauser, U. Between persistently active and activity-silent frameworks: novel vistas on the cellular basis of working memory. *Ann. N. Y. Acad. Sci.* **1464**, 64–75 (2020).
95. Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).
96. O'Connell, R. G., Shadlen, M. N., Wong-Lin, K. & Kelly, S. P. Bridging neural and computational viewpoints on perceptual decision-making. *Trends Neurosci.* **41**, 838–852 (2018).
97. Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* **61**, 331–349 (1989).
98. Constantinidis, C., Franowicz, M. N. & Goldman-Rakic, P. S. Coding specificity in cortical microcircuits: a multiple-electrode analysis of primate prefrontal cortex. *J. Neurosci.* **21**, 3646–3655 (2001).
99. Shadlen, M. N. & Newsome, W. T. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J. Neurophysiol.* **86**, 1916–1936 (2001).
100. Roitman, J. D. & Shadlen, M. N. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci.* **22**, 9475–9489 (2002).
101. Hanks, T. D. et al. Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature* **520**, 220–223 (2015).
102. Goard, M. J., Pho, G. N., Woodson, J. & Sur, M. Distinct roles of visual, parietal, and frontal motor cortices in memory-guided sensorimotor decisions. *eLife* **5**, e13764 (2016).
103. Chandrasekaran, C., Peixoto, D., Newsome, W. T. & Shenoy, K. V. Laminar differences in decision-related neural activity in dorsal premotor cortex. *Nat. Commun.* **8**, 996 (2017).
104. Peixoto, D. et al. Decoding and perturbing decision states in real time. *Nature* **591**, 604–609 (2021).
105. Kilpatrick, Z. P., Ermentrout, B. & Doiron, B. Optimizing working memory with heterogeneity of recurrent cortical excitation. *J. Neurosci.* **33**, 18999–19011 (2013).
106. Barbosa, J. et al. Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nat. Neurosci.* **23**, 1016–1024 (2020).
107. Wimmer, K., Nykamp, D. Q., Constantinidis, C. & Compte, A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat. Neurosci.* **17**, 431–439 (2014).
108. Stein, H. et al. Reduced serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia. *Nat. Commun.* **11**, 4250 (2020).
109. Cano-Colino, M. & Compte, A. A computational model for spatial working memory deficits in schizophrenia. *Pharmacopsychiatry* **45**, S49–S56 (2012).
110. Stein, H., Barbosa, J. & Compte, A. Towards biologically constrained attractor models of schizophrenia. *Curr. Opin. Neurobiol.* **70**, 171–181 (2021).
111. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
112. Chaisangmongkon, W., Swaminathan, S. K., Freedman, D. J. & Wang, X.-J. Computing by robust transience: how the fronto-parietal network performs sequential, category-based decisions. *Neuron* **93**, 1504–1517.e4 (2017).
113. Kobak, D. et al. Demixed principal component analysis of neural population data. *eLife* **5**, e10989 (2016).
114. Aoi, M. C., Mante, V. & Pillow, J. W. Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. *Nat. Neurosci.* **23**, 1410–1420 (2020).
115. Koren, V., Andrei, A. R., Hu, M., Dragoi, V. & Obermayer, K. Reading-out task variables as a low-dimensional reconstruction of neural spike trains in single trials. *PLoS ONE* **14**, e0222649 (2019).
116. Sussillo, D., Churchland, M. M., Kaufman, M. T. & Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* **18**, 1025–1033 (2015).
117. Song, H. F., Yang, G. R. & Wang, X.-J. Training excitatory–inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework. *PLoS Comput. Biol.* **12**, e1004792 (2016).
118. Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T. & Wang, X.-J. Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* **22**, 297–306 (2019).
119. Feulner, B. & Clopath, C. Neural manifold under plasticity in a goal driven learning behaviour. *PLoS Comput. Biol.* **17**, e1008621 (2021).
120. Mastrogiuseppe, F. & Ostojic, S. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron* **99**, 609–623.e29 (2018).
121. Eliasmith, C. & Anderson, C. H. *Neural Engineering (Computational Neuroscience Series): Computational, Representation, and Dynamics in Neurobiological Systems* (MIT Press, 2002).

122. Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F. & Ostojic, S. The role of population structure in computations through neural dynamics. *Nat. Neurosci.* **25**, 783–794 (2022).
123. Valente, A., Ostojic, S. & Pillow, J. Probing the relationship between latent linear dynamical systems and low-rank recurrent neural network models. *Neural Comput.* **34**, 1871–1892 (2022).
124. Langdon, C. & Engel, T. A. Latent circuit inference from heterogeneous neural responses during cognitive tasks. Preprint at *bioRxiv* https://doi.org/10.1101/2022.01.23.477431 (2022).
125. Macke, J. H. et al. in *Advances in Neural Information Processing Systems 24* https://papers.nips.cc/paper_files/paper/2011/hash/7143d7fbadfa4693b9eec507d9d37443-Abstract.html (2011).
126. Gao, Y., Busing, L., Shenoy, K. V. & Cunningham, J. P. in *Advances in Neural Information Processing Systems 28* https://papers.nips.cc/paper_files/paper/2011/hash/7143d7fbadfa4693b9eec507d9d37443-Abstract.html (2015).
127. Rajan, K., Harvey, C. D. & Tank, D. W. Recurrent network models of sequence generation and memory. *Neuron* **90**, 128–142 (2016).
128. Cohen, Z., DePasquale, B., Aoi, M. C. & Pillow, J. W. Recurrent dynamics of prefrontal cortex during context-dependent decision-making. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.27.401539 (2020).
129. Perich, M. G. & Rajan, K. Rethinking brain-wide interactions through multi-region 'network of networks' models. *Curr. Opin. Neurobiol.* **65**, 146–151 (2020).
130. Bittner, S. R. et al. Interrogating theoretical models of neural computation with emergent property inference. *eLife* **10**, e56265 (2021).
131. Friston, K. et al. Dynamic causal modelling revisited. *NeuroImage* **199**, 730–744 (2019).
132. Friston, K. J., Harrison, L. & Penny, W. Dynamic causal modelling. *NeuroImage* **19**, 1273–1302 (2003).
133. Chernov, M. M., Friedman, R. M., Chen, G., Stoner, G. R. & Roe, A. W. Functionally specific optogenetic modulation in primate visual cortex. *Proc. Natl Acad. Sci. USA* **115**, 10505–10510 (2018).
134. Carrillo-Reid, L., Han, S., Yang, W., Akrouh, A. & Yuste, R. Controlling visually guided behavior by holographic recalling of cortical ensembles. *Cell* **178**, 447–457.e5 (2019).
135. Marshel, J. H. et al. Cortical layer-specific critical dynamics triggering perception. *Science* **365**, 6453 (2019).
136. Saxena, S. & Cunningham, J. P. Towards the neural population doctrine. *Curr. Opin. Neurobiol.* **55**, 103–111 (2019).
137. Ebitz, R. B. & Hayden, B. Y. The population doctrine in cognitive neuroscience. *Neuron* **109**, 3055–3068 (2021).
138. Cueva, C. J. et al. Low-dimensional dynamics for working memory and time encoding. *Proc. Natl Acad. Sci. USA* **117**, 23021–23032 (2020).
139. Deneve, S., Alemi, A. & Bourdoukan, R. The brain as an efficient and robust adaptive learner. *Neuron* **94**, 969–977 (2017).
140. Tenenbaum, J. B., Silva, VD & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000).
141. Priebe, N. J. & Ferster, D. Mechanisms of neuronal computation in mammalian visual cortex. *Neuron* **75**, 194–208 (2012).
142. Pollock, E. & Jazayeri, M. Engineering recurrent neural networks from task-relevant manifolds and dynamics. *PLoS Comput. Biol.* **16**, e1008128 (2020).
143. Strang, G. *Introduction to Linear Algebra* (Wellesley-Cambridge, 1998).
144. Kuznetsov, Y. A. *Topological Equivalence, Bifurcations, and Structural Stability of Dynamical Systems* 39–76 (Springer, 2004).

Acknowledgements

This work was supported by the Swartz Foundation (C.L. and M.G.), National Institutes of Health (NIH) grant R01 EB026949 (T.A.E. and M.G.), NIH grant RF1DA055666 (T.A.E., C.L. and M.G.) and Alfred P. Sloan Foundation Research Fellowship (T.A.E.).

Author contributions

The authors contributed equally to all aspects of the article.

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Neuroscience* thanks the anonymous reviewers for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.