Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

HHS Public Access
Author manuscript
Nat Rev Neurosci. Author manuscript; available in PMC 2024 April 29.

Published in final edited form as:
Nat Rev Neurosci. 2023 June ; 24(6): 363–377. doi:10.1038/s41583-023-00693-x.

A unifying perspective on neural manifolds and circuits for cognition

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

[1]Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey, USA

[2]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA

[3]These authors contributed equally

Abstract

Two different perspectives have informed efforts to explain the link between the brain and behavior. 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 centers on neural manifolds — low-dimensional representations of behavioral signals in neural population activity — and suggests that neural computations are realized by emergent dynamics. While 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 underlying behavior.

Introduction

Behavioral and cognitive functions emerge from the dynamic interactions of many neurons wired into circuits. Understanding how circuit connectivity gives rise to neural dynamics and behavior is a central goal in systems neuroscience. This problem, however, remains unresolved, in part due to the experimental challenge of measuring both the connectivity and activity of the same neurons during behavior. While stunning technological advances have enabled us to record activity from increasingly large populations of neurons[1,2,3], the observational data that these experiments generate do not unambiguously point to circuit mechanisms. Likewise, reconstructions of anatomical connectivity[4,5,6,7] constrain the space

* tatiana.engel@princeton.edu .

Author contributions
The authors contributed equally to all aspects of the article.

Competing interests
The authors declare no competing interests.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of possible neural dynamics but do not uniquely predict the activity patterns that arise from the circuit to control specific behaviors. Therefore, theory and computational modeling have been instrumental in bridging the gaps between circuit connectivity, neural dynamics and behavioral functions.

Traditionally, theoretical models hypothesize possible circuit mechanisms to reproduce the neural responses and behavior observed in experiments[8,9,10,11,12,13,14]. In these circuit models, the recurrent connectivity is usually hand-crafted to produce the neural activity patterns that are needed to solve a particular behavioral 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 stables states of the system) in neural dynamics, which can support categorical decision-making[10,15,16,17]. 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 behavioral 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[23,24,25].

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[27,28,29,30]. Tying this complex and heterogeneous activity to the underlying circuit mechanism thus poses a formidable challenge.

Over the last decade, multiple statistical techniques have emerged that can find structure in heterogeneous neural responses[31,32,33,34] (Box 1). Although diverse, the responses of different neurons in a population are usually tightly correlated during behavioral 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 behavioral task execution. Modeling how neural population dynamics unfold along the manifold as the task progresses provides a dynamical-system description of neural computation[35,36,37,38,39]. The discovery of interpretable manifolds in multiple brain areas and behavioral 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].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

While neural manifolds concisely summarize heterogeneous single-cell responses, they only provide 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. While several recent reviews have highlighted insights provided by studies focusing on neural manifolds[32,40,41,42] and circuits[43,44] separately, we here advocate for the integration of neural manifold and circuit approaches to cognition. We review the fly's head direction (HD) 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 behavior.

Circuit–manifold convergence: HD 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[45,46,47]. The function of the HD system is to represent the direction that 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 HD 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 due to the simple physical layout of the ellipsoid body (EB), the core neuropil in the fly HD system[49], within which topographically organized neural responses can be directly visualized using calcium imaging (Fig. 1b). The EB has a circular structure and the neurons within the EB (called E-PG neurons) display a 'bump' of activity within this circle, indicating that a small set of neighboring neurons are 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 HD system is more complex, involving many brain regions in which HD cells are scattered, and there is no evidence so far for a topographic ring structure organization[46,53]. Nevertheless, the responses of HD cells in the anterodorsal thalamic nucleus[48] and postsubiculum[54] of mice also form a ring manifold. The ring is non-linearly 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, multi-modal 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 HD 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, due 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 due to local recurrent excitation and unique due 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 HD cells[57] (Fig. 1c), while 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 HD cells. Each group of rotation cells receives input from its corresponding group of HD cells and projects to the clockwise or anticlockwise neighbors of that HD cell (Fig. 1c).

The ring attractor model accounts for all salient features of the ring manifold and dynamics in the fly EB, although it does not reproduce the complex tuning curves of some HD 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 HD 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 HD cells rather than local recurrent

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

excitation[46,58], consistent with the lack of strong recurrent connections between HD cells that has been observed in 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 *Drosophila 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 HD system, allowing for the ring attractor model to be tested directly. It was found that local excitation between E-PG neurons sustains the activity bump in the EB. The rotation cells predicted by the model were identified as P-EN neurons located in another neuropil called the protocerebral bridge (PB). 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 PB 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[50,57,62,63,64].

The fruit fly HD 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 behavior. While neural responses in the rodent HD system organize on a similar ring manifold, they show more heterogeneity and lack a clear spatial topography. The precise circuit mechanism generating dynamics on the ring manifold in the rodent HD 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

position coincides with any vertex of a regular grid of equilateral triangles spanning the environment. Grids of neighboring 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 HD 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. Since the tuning of single cells within a module varies only by spatial phase, it is expected that their population responses form a 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 behavioral 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 preferentially fired 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 HD 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 is unknown. For example, during sleep, estimating single-cell tuning curves is not possible because there are no behavioral variables to which neural activity can be

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

referenced. However, manifold analyses reveal the same manifolds in HD cells[48,54] and grid cells[71] during sleep as during wakefulness, suggesting that they arise from the anatomical connectivity. Similarly, manifold analyses enabled the discovery of a HD circuit in the anterior hindbrain of larval zebrafish, by demonstrating that neural responses in this area form a ring manifold[73]. Since 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 orderly encoded. The manifold thus formed a conjoined cognitive map of the task[74], with some dimensions reflecting information beyond the measured behavioral variables[75]. While the activity of single place cells in 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 HD 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. 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 due to 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[79,80,81], but this mechanism is inconsistent with the rigidity of the toroidal manifold across environments and behavioral states. By contrast, the spatially arranged connectivity in attractor models naturally leads to a rigid manifold structure that does not change with varying input[77,78,82].

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[76,82], but which one applies to the grid-cell system is unknown. Unlike the fly HD 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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[83] and drug infusions to alter the gain of recurrent inhibition[84] 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 lack simple 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 HD cells and grid cells[59,85,86]. 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 (as is also the case in the rodent's navigational circuits). 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[59,86] (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 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[59,86] (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: while measuring the activity and anatomical connectivity of the same neurons can enable direct testing of hypothesized circuit mechanisms with single-cell resolution[86,87], 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 the higher cortical areas supporting cognitive functions, such as working memory[88,89,90,91] or decision making[92,93]. Tasks used to study cognitive functions in animal experiments often have a simple topological structure (Box 1), akin

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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[94,95] (Fig. 3a). The remembered angle is a one-dimensional circular variable, just like 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[92]. 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 spatial tuning, providing an essential neural correlate of working memory[88,90,91,94,95]. 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[92,96,97,98,99,100,101].

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 HD cells) captures the stimulus-dependent persistent activity observed during spatial working memory tasks[9,102,103] (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[104] as well as memory deficits arising from circuit disruptions (such as altered excitation-inhibition balance) in mental illness[21,105,106,107]. The ramping activity associated with decision-making arises in discrete attractor models[10,15,16,17,18]. 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 behavior that occur when optogenetic perturbations of neural activity are performed in rodents[23,24,25].

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[27,28,29,30,108,109]. Only a relatively small fraction (5-10%) of PFC neurons shows 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[108,109]. 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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[31,110,111,112]. 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[108,109] (Fig. 3c).

Manifold analyses have uncovered manifolds mirroring the topology of task variables in many cortical areas and cognitive tasks[27,28,29,30,108,109]. Manifolds and single-neuron heterogeneity qualitatively similar to that observed in brain recordings also emerge in RNNs trained to perform cognitive tasks by optimizing recurrent connectivity parameters[108,109,113,114,115,116] (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?

Linking manifolds to connectivity

Low-rank recurrent neural networks.—It is possible to reconcile the manifold and circuit perspectives by engineering non-linear 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[117,118]. 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). 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[117,119]. It is, however, important to note that, although low-rank RNNs can be engineered to solve certain tasks, it does not

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 non-linear 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[120]. Linear dynamical systems provide a mathematically tractable approximation to non-linear brain dynamics[108]. 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 (Box 4)[120]. 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[119,121]. 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[119,121].

To what extent does the intuition gained from such linear systems extend to non-linear recurrent networks such as RNNs or the brain? A recent preprint[121] 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 non-linear 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 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

structure in these networks, which was validated by patterned perturbations of the RNN activity and connectivity[121]. 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[122,123], but incorporates a biologically plausible non-linearity 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[124,125,126]. However, the full high-dimensional connectivity is not uniquely constrained by low-dimensional neural trajectories and is not easily interpretable[86] (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[127,128,129] 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 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, principle component analysis looks for the projections that account for the most variance in the response data, irrespective of whether or not 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[30,108,110,111]. 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[30,108,110,111]. 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[121]. These results thus highlight the importance of interpreting neural representations within the context of circuit mechanisms.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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[86,117,119,121]. 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 behavior. Experiments can test these causal relationships by validating the behavioral 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[130,131,132]. While 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 HD 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[133,134]. 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 behavioral 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[49,68,86,117,119,121]. 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 confirmed 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[86,87,135]. 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

can be related to a low-dimensional connectivity structure that gives rise to the manifold in neural activity space[117,119,121]. This relationship was directly confirmed in RNNs[121], and future work can test it in biological data.

Similar to the classical circuit models, the neural manifold and 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[117,118,119,121,136]. 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 modeling that is needed to identify the circuit mechanisms giving rise to the neural dynamics that drive behavior.

Acknowledgements

This work was supported by the Swartz Foundation (C.L. and M.G.), the 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.).

References

1. Steinmetz NA, Koch C, Harris KD & Carandini M Challenges and opportunities for large-scale electrophysiology with Neuropixels probes. Current Opinion in Neurobiology 50, 92–100 (2018). [PubMed: 29444488]
2. Steinmetz NA et al. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. Science 372, eabf4588 (2021). [PubMed: 33859006]
3. Ebrahimi S. et al. Emergent reliability in sensory cortical coding and inter-area communication. Nature 605, 713–721 (2022). [PubMed: 35589841]
4. Oh SW et al. A mesoscale connectome of the mouse brain. Nature 508, 207–214 (2014). [PubMed: 24695228]
5. Markov NT et al. Cortical High-Density Counterstream Architectures. Science (New York, N.Y.) 342, 1238406–1238406 (2013). [PubMed: 24179228]
6. Harris JA et al. Hierarchical organization of cortical and thalamic connectivity. Nature 508, 207–30 (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). [PubMed: 32619423]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

8. Skaggs WE, Knierim JJ, Kudrimoti HS & McNaughton BL A model of the neural basis of the rat's sense of direction. Advances in neural information processing systems 7, 173–80 (1995). [PubMed: 11539168]
9. Compte A, Brunel N, Goldman-Rakic PS & Wang X-J Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cerebral cortex 10, 910–923 (2000). [PubMed: 10982751]
10. Wang XJ Probabilistic decision making by slow reverberation in cortical circuits. Neuron 36, 955–968 (2002). [PubMed: 12467598]
11. Machens CK, Romo R & Brody CD Flexible Control of Mutual Inhibition: A Neural Model of Two-Interval Discrimination. Science (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). [PubMed: 16767089]
13. Engel TA & Wang XJ Same or different? A neural circuit mechanism of similarity-based pattern match decision making. Journal of Neuroscience 31, 6982–6996 (2011). [PubMed: 21562260]
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, 19504–17 (2013). [PubMed: 24336717]
15. Roach JP, Churchland AK & Engel TA Two roles for choice selective inhibition in decision-making circuits. bioRxiv 2022.01.24.477635 (2022).
16. Martí D, Deco G, Mattia M, Gigante G & Giudice PD A Fluctuation-Driven Mechanism for Slow Decision Processes in Reverberant Networks. PLoS ONE 3, e2534 (2008). [PubMed: 18596965]
17. Ksander J, Katz DB & Miller P A model of naturalistic decision making in preference tests. PLoS Computational Biology 17, e1009012 (2021). [PubMed: 34555012]
18. Wong K-F & Wang X-J A recurrent network mechanism of time integration in perceptual decisions. J Neurosci 26, 1314–1328 (2006). [PubMed: 16436619]
19. Wang X-J The Prefrontal Cortex as a Quintessential "Cognitive-Type" Neural Circuit. Principles of Frontal Lobe Function (Principles of Frontal Lobe Function, 2013).
20. McNaughton BL, Battaglia FP, Jensen O, Moser EI & Moser M-B Path integration and the neural basis of the 'cognitive map'. Nature Reviews Neuroscience 7, 663–678 (2006). [PubMed: 16858394]
21. Murray JD et al. Linking microcircuit dysfunction to cognitive impairment: effects of dis-inhibition associated with schizophrenia in a cortical working memory model. Cerebral cortex 24, 859–872 (2014). [PubMed: 23203979]
22. Lam NH et al. Effects of Altered Excitation-Inhibition Balance on Decision Making in a Cortical Circuit Model. The Journal of Neuroscience JN–RM–1371–20 (2021).
23. Inagaki HK, Fontolan L, Romani S & Svoboda K Discrete attractor dynamics underlies persistent activity in the frontal cortex. Nature 566, 212–217 (2019). [PubMed: 30728503]
24. Finkelstein A. et al. Attractor dynamics gate cortical information flow during decision-making. Nature Neuroscience 24, 843–850 (2021). [PubMed: 33875892]
25. Duan CA et al. Collicular circuits for flexible sensorimotor routing. Nature Neuroscience 24, 1110–1120 (2021). [PubMed: 34083787]
26. Urai AE, Doiron B, Leifer AM & Churchland AK Large-scale neural recordings call for new insights to link brain and behavior. Nature Neuroscience 1–9 (2022).
27. Fusi S, Miller EK & Rigotti M Why neurons mix: high dimensionality for higher cognition. Current Opinion in Neurobiology 37, 66–74 (2016). [PubMed: 26851755]
28. Cavanagh SE, Towers JP, Wallis JD, Hunt LT & Kennerley SW Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. Nature Communications 9, 3498 (2018).
29. Murray JD et al. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. Proc Natl Acad Sci U S A 114, 394–399 (2017). [PubMed: 28028221]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

30. Machens CK, Romo R & Brody CD Functional, But Not Anatomical, Separation of ”What” and ”When” in Prefrontal Cortex. Journal of Neuroscience 30, 350–360 (2010). [PubMed: 20053916]
31. Cunningham JP & Yu BM Dimensionality reduction for large-scale neural recordings. Nat. Neurosci 17, 1500–1509 (2014). [PubMed: 25151264]
32. Gallego JA, Perich MG, Miller LE & Solla SA Neural manifolds for the control of movement. Neuron 94, 978–984 (2017). [PubMed: 28595054]
33. Duncker L & Sahani M Dynamics on the manifold: Identifying computational dynamical activity from neural population recordings. Current Opinion in Neurobiology 70, 163–170 (2021). [PubMed: 34837752]
34. Kriegeskorte N & Wei X-X Neural tuning and representational geometry. Nature Reviews Neuroscience 22, 703–718 (2021). [PubMed: 34522043]
35. Pandarinath C. et al. Inferring single-trial neural population dynamics using sequential autoencoders. Nature methods 15, 805–815 (2018). [PubMed: 30224673]
36. Duncker L, Bohner G, Boussard J & Sahani M Learning interpretable continuous-time models of latent stochastic dynamical systems. arXiv preprint at https://arxiv.org/abs/1902.04420 (2019).
37. Genkin M & Engel TA Moving beyond generalization to accurate interpretation of flexible models. Nature Machine Intelligence 2, 674–683 (2020).
38. Zhao Y & Park IM Variational Online Learning of Neural Dynamics. Frontiers in Computational Neuroscience 14, 71 (2020). [PubMed: 33154718]
39. Genkin M, Hughes O & Engel TA Learning non-stationary Langevin dynamics from stochastic observations of latent trajectories. Nature Communications 12, 5986 (2021).
40. Vyas S, Golub MD, Sussillo D & Shenoy KV Computation Through Neural Population Dynamics. Annual Review of Neuroscience 43, 249–275 (2020).
41. Barack DL & Krakauer JW Two views on the cognitive brain. Nature Reviews Neuroscience 22, 359–371 (2021). [PubMed: 33859408]
42. Jazayeri M & Ostojic S Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. Current opinion in neurobiology 70, 113–120 (2021). [PubMed: 34537579]
43. Knierim JJ & Zhang K Attractor Dynamics of Spatially Correlated Neural Activity in the Limbic System. Annual review of neuroscience 35, 267–285 (2012).
44. Khona M & Fiete IR Attractor and integrator networks in the brain. arXiv (2021). 2112.03978.
45. Clark BJ & Taube JS Vestibular and attractor network basis of the head direction cell signal in subcortical circuits. Frontiers in Neural Circuits 6, 7 (2012). [PubMed: 22454618]
46. Hulse BK & Jayaraman V Mechanisms Underlying the Neural Computation of Head Direction. Annual Review of Neuroscience 43, 1–24 (2016).
47. Turner-Evans DB et al. The neuroanatomical ultrastructure and function of a biological ring attractor. Neuron 108, 145–163 (2020). [PubMed: 32916090]
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). [PubMed: 31406365]
49. Seelig JD & Jayaraman V Neural dynamics for landmark orientation and angular path integration. Nature 521, 186–191 (2015). [PubMed: 25971509]
50. Turner-Evans D. et al. Angular velocity integration in a fly heading circuit. Elife 6, e23496 (2017). [PubMed: 28530551]
51. Kim SS, Rouault H, Druckmann S & Jayaraman V Ring attractor dynamics in the Drosophila central brain. Science (2017).
52. Kim SS, Hermundstad AM, Romani S, Abbott LF & Jayaraman V Generation of stable heading representations in diverse visual scenes. Nature 576, 126–131 (2019). [PubMed: 31748750]
53. Ajabi Z, Keinath AT, Wei X-X & Brandon MP Population dynamics of the thalamic head direction system during drift and reorientation. bioRxiv 2021.08.30.458266 (2021).
54. Duszkiewicz AJ et al. Reciprocal feature encoding by cortical excitatory and inhibitory neurons. bioRxiv 2022.03.14.484357 (2022).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

55. Zhang K. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. The Journal of Neuroscience 16, 2112–2126 (1996). [PubMed: 8604055]
56. Redish AD, Elga AN & Touretzky DS A coupled attractor model of the rodent head direction system. Network: Computation in Neural Systems 7, 671–685 (1996).
57. Cope AJ, Sabo C, Vasilaki E, Barron AB & Marshall JAR A computational model of the integration of landmarks and motion in the insect central complex. PLoS ONE 12, e0172325 (2017). [PubMed: 28241061]
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). [PubMed: 15673682]
59. Cueva CJ, Wang PY, Chin M & Wei X-X Emergence of functional and structural properties of the head direction system by optimization of recurrent neural networks. arXiv (2019). 1912.10189.
60. Green J. et al. A neural circuit architecture for angular integration in Drosophila. Nature 546, 101–106 (2017). [PubMed: 28538731]
61. Wolff T, Iyer NA & Rubin GM Neuroarchitecture and neuroanatomy of the Drosophila central complex: A GAL4-based dissection of protocerebral bridge neurons and circuits. Journal of Comparative Neurology 523, 997–1037 (2015). [PubMed: 25380328]
62. Kutschireiter A, Basnak MA, Wilson RI & Drugowitsch J A Bayesian perspective on the ring attractor for heading-direction tracking in the Drosophila central complex. bioRxiv 2021.12.17.473253 (2021).
63. Lyu C, Abbott LF & Maimon G Building an allocentric travelling direction signal via vector computation. Nature 1–6 (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. Nature Communications 8, 139 (2017).
65. Mittelstaedt ML & Mittelstaedt H Homing by path integration in a mammal. Naturwissenschaften 67, 566–567 (1980).
66. Fyhn M, Molden S, Witter MP, Moser EI & Moser M-B Spatial representation in the entorhinal cortex. Science 305, 1258–1264 (2004). [PubMed: 15333832]
67. Stensola H. et al. The entorhinal grid map is discretized. Nature 492, 72–78 (2012). [PubMed: 23222610]
68. Yoon K. et al. Specific evidence of low-dimensional continuous attractor dynamics in grid cells. Nat. Neurosci 16, 1077–1084 (2013). [PubMed: 23852111]
69. Hafting T, Fyhn M, Molden S, Moser M-B & Moser EI Microstructure of a spatial map in the entorhinal cortex. Nature 436, 801–806 (2005). [PubMed: 15965463]
70. Gao R, Xie J, Wei X-X, Zhu S-C & Wu YN On Path Integration of Grid Cells: Group Representation and Isotropic Scaling. arXiv (2020). 2006.10259.
71. Gardner RJ et al. Toroidal topology of population activity in grid cells. Nature 602, 123–128 (2022). [PubMed: 35022611]
72. Fiete IR, Burak Y & Brookings T What grid cells convey about rat location. Journal of Neuroscience 28, 6858–6871 (2008). [PubMed: 18596161]
73. Petrucco L. et al. Neural dynamics and architecture of the heading direction circuit in a vertebrate brain. bioRxiv 2022.04.27.489672 (2022).
74. Nieh EH et al. Geometry of abstract learned knowledge in the hippocampus. Nature 1–5 (2021).
75. Low RJ, Lewallen S, Aronov D, Nevers R & Tank DW Probing variability in a cognitive map using manifold inference from neural dynamics. bioRxiv 418939 (2018).
76. Widloski J, Marder MP & Fiete IR Inferring circuit mechanisms from sparse neural recording and global perturbation in grid cells. eLife 7, e33503 (2018). [PubMed: 29985132]
77. Samsonovich A & McNaughton BL Path integration and cognitive mapping in a continuous attractor neural network model. The Journal of Neuroscience 17, 5900–5920 (1997). [PubMed: 9221787]
78. Conklin J & Eliasmith C A controlled attractor network model of path integration in the rat. Journal of Computational Neuroscience 18, 183–203 (2005). [PubMed: 15714269]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

79. Kropff E & Treves A The emergence of grid cells: Intelligent design or just adaptation? Hippocampus 18, 1256–1269 (2008). [PubMed: 19021261]
80. 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). [PubMed: 26952211]
81. Monsalve-Mercado MM & Leibold C Hippocampal spike-timing correlations lead to hexagonal grid Fields. Physical Review Letters 119, 038101 (2017). [PubMed: 28777606]
82. Burak Y & Fiete IR Accurate path integration in continuous attractor network models of grid cells. PLoS Computational Biology 5, e1000291 (2009). [PubMed: 19229307]
83. Banerjee A, Egger R & Long MA Using focal cooling to link neural dynamics and behavior. Neuron 109, 2508–2518 (2021). [PubMed: 34171292]
84. Rudolph U & Möhler H Analysis of GABAA receptor function and dissection of the pharmacology of benzodiazepines and general anesthetics through mouse genetics . Annual review of pharmacology and toxicology 44, 475–498 (2004).
85. Cueva CJ & Wei X-X Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. arXiv (2018). 1803.07770.
86. Sorscher B, Mel GC, Ocko SA, Giocomo LM & Ganguli S A unified theory for the computational and mechanistic origins of grid cells. Neuron (2022).
87. Bock DD et al. Network anatomy and in vivo physiology of visual cortical neurons. Nature 471, 177–182 (2011). [PubMed: 21390124]
88. Wang X-J 50 years of mnemonic persistent activity: quo vadis? Trends in Neurosciences 44, 888–902 (2021). [PubMed: 34654556]
89. Fuster JM & Alexander GE Neuron Activity Related to Short-Term Memory. Science 173, 652–654(1971). [PubMed: 4998337]
90. Romo R, Brody CD, Hernández A & Lemus L Neuronal correlates of parametric working memory in the prefrontal cortex. Nature 399, 470–473 (1999). [PubMed: 10365959]
91. Kamiński J & Rutishauser U Between persistently active and activity-silent frameworks: novel vistas on the cellular basis of working memory. Annals of the New York Academy of Sciences 1464, 64–75 (2020). [PubMed: 31407811]
92. Gold JI & Shadlen MN The neural basis of decision making. Annu Rev Neurosci 30, 535–574 (2007). [PubMed: 17600525]
93. O'Connell RG, Shadlen MN, Wong-Lin K & Kelly SP Bridging Neural and Computational Viewpoints on Perceptual Decision-Making. Trends in Neurosciences 41, 838–852 (2018). [PubMed: 30007746]
94. Funahashi S, Bruce CJ & Goldman-Rakic PS Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. Journal of Neurophysiology 61, 331–349 (1989). [PubMed: 2918358]
95. Constantinidis C, Franowicz MN & Goldman-Rakic PS Coding Specificity in Cortical Microcircuits: A Multiple-Electrode Analysis of Primate Prefrontal Cortex. The Journal of Neuroscience 21, 3646–3655 (2001). [PubMed: 11331394]
96. Shadlen MN & Newsome WT Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. Journal of Neurophysiology 86, 1916–1936 (2001). [PubMed: 11600651]
97. Roitman JD & Shadlen MN Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. J Neurosci 22, 9475–9489 (2002). [PubMed: 12417672]
98. Hanks TD et al. Distinct relationships of parietal and prefrontal cortices to evidence accumulation. Nature 520, 220–223 (2015). [PubMed: 25600270]
99. Goard MJ, Pho GN, Woodson J & Sur M Distinct roles of visual, parietal, and frontal motor cortices in memory-guided sensorimotor decisions. eLife 5, e13764 (2016). [PubMed: 27490481]
100. Chandrasekaran C, Peixoto D, Newsome WT & Shenoy KV Laminar differences in decision-related neural activity in dorsal premotor cortex. Nature Communications 8, 996 (2017).
101. Peixoto D. et al. Decoding and perturbing decision states in real time. Nature 591, 604–609 (2021). [PubMed: 33473215]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

102. Kilpatrick ZP, Ermentrout B & Doiron B Optimizing working memory with heterogeneity of recurrent cortical excitation. Journal of Neuroscience 33, 18999–19011 (2013). [PubMed: 24285904]
103. Barbosa J. et al. Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. Nature Neuroscience 23, 1016–1024 (2020). [PubMed: 32572236]
104. Wimmer K, Nykamp DQ, Constantinidis C & Compte A Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nat Neurosci 17, 431–439 (2014). [PubMed: 24487232]
105. Stein H. et al. Reduced serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia. Nature Communications 11, 4250 (2020).
106. Cano-Colino M & Compte A A Computational Model for Spatial Working Memory Deficits in Schizophrenia. Pharmacopsychiatry 45, S49–S56 (2012). [PubMed: 22565235]
107. Stein H, Barbosa J & Compte A Towards biologically constrained attractor models of schizophrenia. Current Opinion in Neurobiology 70, 171–181 (2021). [PubMed: 34839146]
108. Mante V, Sussillo D, Shenoy KV & Newsome WT Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature 503, 78–84 (2013). [PubMed: 24201281]
109. Chaisangmongkon W, Swaminathan SK, Freedman DJ & Wang X-J Computing by Robust Transience: How the Fronto-Parietal Network Performs Sequential, Category-Based Decisions. Neuron 93, 1504–1517.e4 (2017). [PubMed: 28334612]
110. Kobak D. et al. Demixed principal component analysis of neural population data. eLife 5, e10989 (2016). [PubMed: 27067378]
111. Aoi MC, Mante V & Pillow JW Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. Nature Neuroscience 23, 1410–1420 (2020). [PubMed: 33020653]
112. Koren V, Andrei AR, 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). [PubMed: 31622346]
113. Sussillo D, Churchland MM, Kaufman MT & Shenoy KV A neural network that finds a naturalistic solution for the production of muscle activity. Nature Neuroscience 18, 1025–1033 (2015). [PubMed: 26075643]
114. Song HF, Yang GR & Wang X-J Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework. PLOS Computational Biology 12, e1004792 (2016). [PubMed: 26928718]
115. Yang GR, Joglekar MR, Song HF, Newsome WT & Wang X-J Task representations in neural networks trained to perform many cognitive tasks. Nature neuroscience 22, 297–306 (2019). [PubMed: 30643294]
116. Feulner B & Clopath C Neural manifold under plasticity in a goal driven learning behaviour. PLoS Computational Biology 17, e1008621 (2021). [PubMed: 33544700]
117. Mastrogiuseppe F & Ostojic S Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks. Neuron 99, 609–623.e29 (2018). [PubMed: 30057201]
118. Eliasmith C & Anderson CH Neural Engineering (Computational Neuroscience Series): Computational, Representation, and Dynamics in Neurobiological Systems (MIT Press, Cambridge, MA, USA, 2002).
119. Dubreuil A, Valente A, Beiran M, Mastrogiuseppe F & Ostojic S The role of population structure in computations through neural dynamics. Nature Neuroscience 25, 783–794 (2022). [PubMed: 35668174]
120. Valente A, Ostojic S & Pillow J Probing the Relationship Between Latent Linear Dynamical Systems and Low-Rank Recurrent Neural Network Models. Neural computation 1–22 (2021).
121. Langdon C & Engel TA Latent circuit inference from heterogeneous neural responses during cognitive tasks. bioRxiv 2022.01.23.477431 (2022).
122. Macke JH et al. Empirical models of spiking in neural populations. In Shawe-Taylor J, Zemel R, Bartlett P, Pereira F & Weinberger K (eds.) Advances in Neural Information Processing Systems, vol. 24 (2011).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

123. Gao Y, Busing L, Shenoy KV & Cunningham JP High-dimensional neural spike train analysis with generalized count linear dynamical systems. In Cortes C, Lawrence N, Lee D, Sugiyama M & Garnett R (eds.) Advances in Neural Information Processing Systems, vol. 28 (2015).
124. Rajan K, Harvey CD & Tank DW Recurrent Network Models of Sequence Generation and Memory. Neuron 90, 128–142 (2016). [PubMed: 26971945]
125. Cohen Z, DePasquale B, Aoi MC & Pillow JW Recurrent dynamics of prefrontal cortex during context-dependent decision-making. bioRxiv 2020.11.27.401539 (2020).
126. Perich MG & Rajan K Rethinking brain-wide interactions through multi-region 'network of networks' models. Current Opinion in Neurobiology 65, 146–151 (2020). [PubMed: 33254073]
127. Bittner SR et al. Interrogating theoretical models of neural computation with emergent property inference. eLife 10, e56265 (2021). [PubMed: 34323690]
128. Friston K. et al. Dynamic causal modelling revisited. NeuroImage 199, 730–744 (2019). [PubMed: 28219774]
129. Friston KJ, Harrison L & Penny W Dynamic causal modelling. NeuroImage 19, 1273–1302 (2003). [PubMed: 12948688]
130. Chernov MM, Friedman RM, Chen G, Stoner GR & Roe AW Functionally specific optogenetic modulation in primate visual cortex. Proceedings of the National Academy of Sciences of the United States of America 115, 10505–10510 (2018). [PubMed: 30257948]
131. 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). [PubMed: 31257030]
132. Marshel JH et al. Cortical layer–specific critical dynamics triggering perception. Science 365 (2019).
133. Saxena S & Cunningham JP Towards the neural population doctrine. Current Opinion in Neurobiology 55, 103–111 (2019). [PubMed: 30877963]
134. Ebitz RB & Hayden BY The population doctrine in cognitive neuroscience. Neuron 109, 3055–3068 (2021). [PubMed: 34416170]
135. Cueva CJ et al. Low-dimensional dynamics for working memory and time encoding. Proceedings of the National Academy of Sciences 117, 23021–23032 (2020).
136. Deneve S, Alemi A & Bourdoukan R The Brain as an Efficient and Robust Adaptive Learner. Neuron 94, 969–977 (2017). [PubMed: 28595053]
137. Tenenbaum JB, Silva V. d. & Langford JC A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290, 2319–2323 (2000). [PubMed: 11125149]
138. Priebe NJ & Ferster D Mechanisms of Neuronal Computation in Mammalian Visual Cortex. Neuron 75, 194–208 (2012). [PubMed: 22841306]
139. Pollock E & Jazayeri M Engineering recurrent neural networks from task-relevant manifolds and dynamics. PLoS Computational Biology 16, e1008128 (2020). [PubMed: 32785228]
140. Strang G. Introduction to Linear Algebra (Wellesley, MA: Wellesley-Cambridge Press, 1998).
141. Kuznetsov YA Topological Equivalence, Bifurcations, and Structural Stability of Dynamical Systems, 39–76 (Springer New York, New York, NY, 2004). URL 10.1007/978-1-4757-3978-7_2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Box 1:

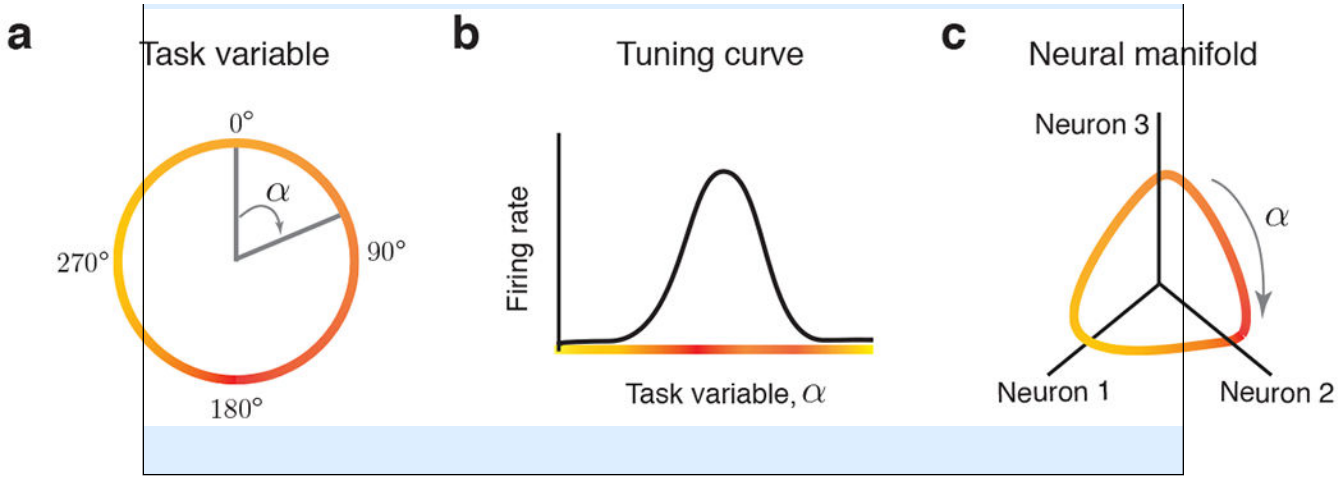
Neural manifolds

Task variables refer to the discrete or continuous parameters of a task, as well as related variables for intermediate representations and computations. These include stimulus parameters (such as the orientation or color 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 the 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 centered at the preferred orientation angle of the neuron[138]. In many brain regions, the responses of single neurons are complex and heterogeneous, with each neuron being tuned for a mixture of task variables[27,28,29,30,48] 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[31,32,33,34,42]. 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 non-linearly embedded in the high-dimensional neural population state space. The shape of this embedding (i.e. the manifold's geometry) depends on the tuning curves of all neurons in the population[34].

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript



Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Box 2:

Neural circuits

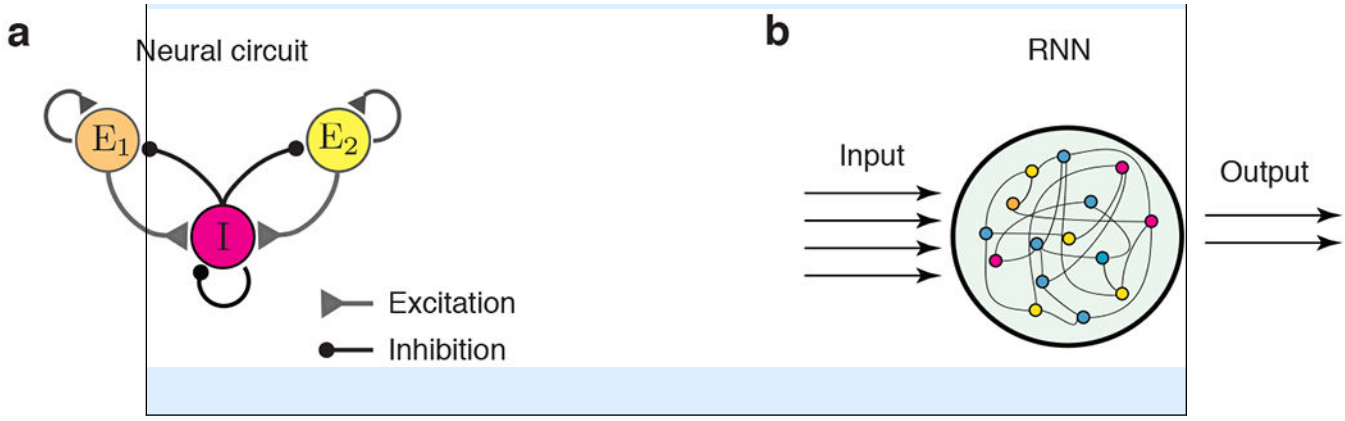
A neural circuit is a network of nodes in which the connections, dynamics of each node's activity and inputs are specified. According to the traditional view, the nodes in the network can be individual neurons, clusters of neurons or brain areas. The connections between nodes implement neural computations. 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 the head direction, see Fig. 1b,c).

Traditional neural circuit models implement interpretable mechanisms that can relate connectivity to dynamics and behavior. Interpretable mechanisms enable us to predict the behavioral effects of specific circuit perturbations, making it possible to experimentally test circuit mechanisms[21,22,23,24,25]. However, the simple connectivity structure of these traditional circuit models results in homogeneous tuning (e.g., 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. A 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 behavioral responses in a cognitive task[108,109,113,114,115,116], low-dimensional manifolds[139] or recorded brain activity (in which case the activity of each RNN unit tracks a target experimental neuron[124]). Distributed mixed selectivity emerges in RNNs through training[108]. 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[114]. 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[117,118,121,136], 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[86,117], making it challenging to identify. However, traditional circuit models can guide the search for interpretable mechanisms in RNNs[86,135].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 non-linear 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)[140]. 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):

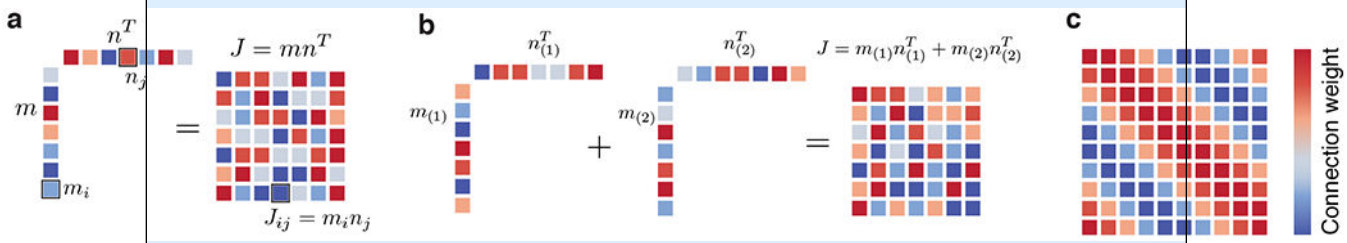
$$J = mn^T, \; J_{ij} = m_i n_j .$$

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 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[140]. 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 article, we highlight examples in which it has been possible to establish a relationship between the low-dimensional activity and connectivity[55,76,86,117,121]. Finding such cases is important because they reveal interpretable circuit mechanisms that can be validated in perturbation experiments.



Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 a $n \times n$ 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 $N \times n (n \ll N)$[140] (see the figure, part **b**). The dynamics of y are then described by a linear dynamical system

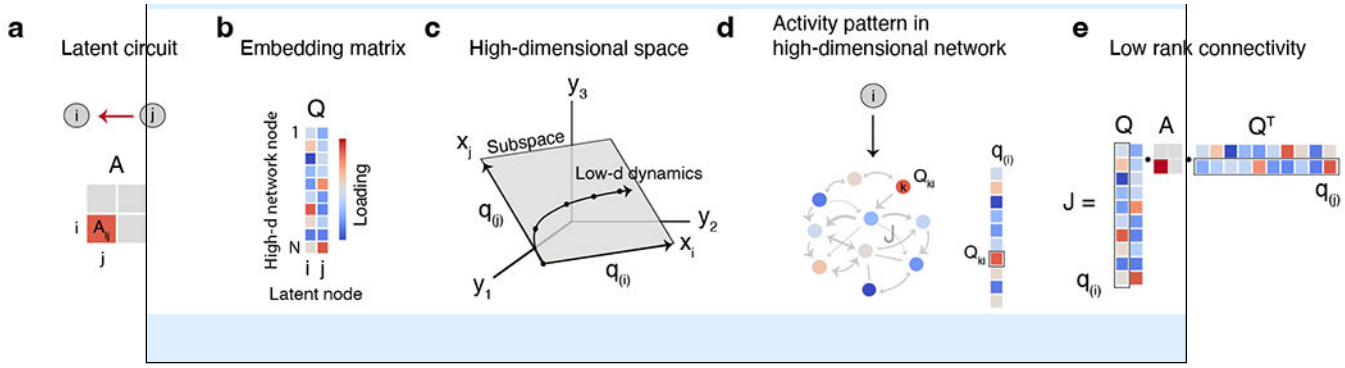
$$\dot{y} = Jy \tag{2}$$

with the low-rank connectivity matrix $J = QAQ^T$ of size $N \times N$[141]. 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 i-th column of Q (see the figure, part **d**). Moreover, writing J as a sum of low-rank 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[121].

Recent theoretical work suggests that low-dimensional dynamics can arise from low-rank connectivity in non-linear RNNs[117,118,119,121,136]. Under what conditions low-dimensional manifolds arise from low-dimensional connectivity in non-linear recurrent networks is an open question. Finding such cases will enable validation of distributed circuit mechanisms via perturbations[121].



Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

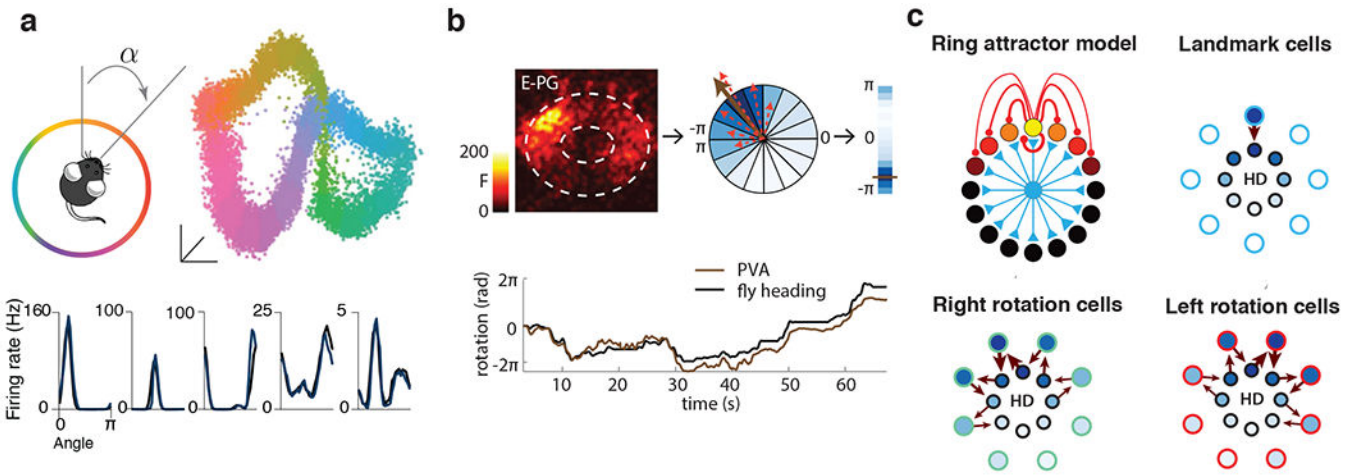


Figure 1. Convergence of a manifold and circuit in the head direction system.

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[137] (upper right). Each dot corresponds to the population activity state at a single time. Color code 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 *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 is fluorescence intensity in arbitrary units)[49,50]. Upper center: The population vector average (PVA, gold arrow) estimates the position of the center of the bump on the circle by summing vectors (dashed red arrows) pointing in the directions of each of 22.5° wedges around the EB with length equal to the instantaneous calcium activity in each wedge. Shade of blue indicates calcium activity in each wedge. EB is unwrapped in a vertical axis (upper right, gold bar indicates angle of PVA) to display 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 (HD) cells[55,56]. The model consists of HD cells arranged in a ring and receiving local excitatory (red) and uniform inhibitory (blue) connections (upper left, cells are colored according to their preferred head direction). Landmark cells provide direct input to the corresponding HD cells (upper right, shade of blue represents the firing rate). Left and right rotation cells (lower panels) make asymmetric recurrent connections with HD cells, projecting strongly to either the left or right neighbor of the HD cell from which they receive input. Upper right and lower images in panel a are adapted, with permission, from Ref.[48] Panel b is adapted, with permission, from Ref.[50] Panel c is adapted, with permission from Refs.[46,51]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

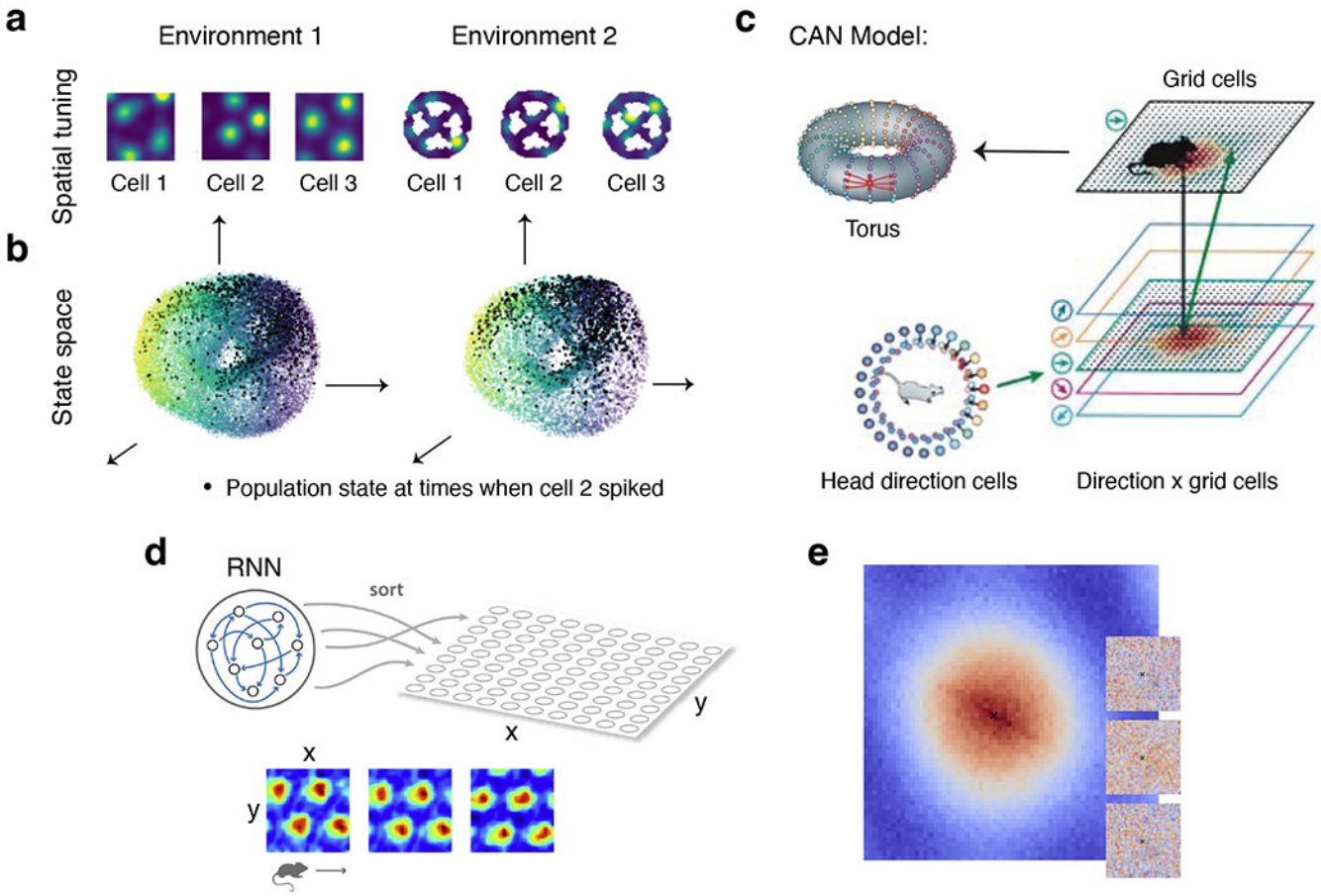


Figure 2. Manifolds and circuits for spatial position encoding.

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 are colored 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 (CAN) 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 the 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 the 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,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 RNNs trained to perform path integration[85,86]. Initially, RNN units are not arranged in any space. After training, RNN units can be sorted on a two-dimensional sheet (upper panel) so that units with similar phases of their grid tuning are close in space. Stable activity patterns on this neural sheet (lower panel, shown at three spacial locations) reveal a topographic hexagonal grid activation as in classical attractor models. **e**, The connections from an individual RNN unit to its neighbors 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, with permission from Ref.[71]. Part c is adapted, with permission, from Ref.[20]. Parts d and e are adapted, with permission, from Ref.[86]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

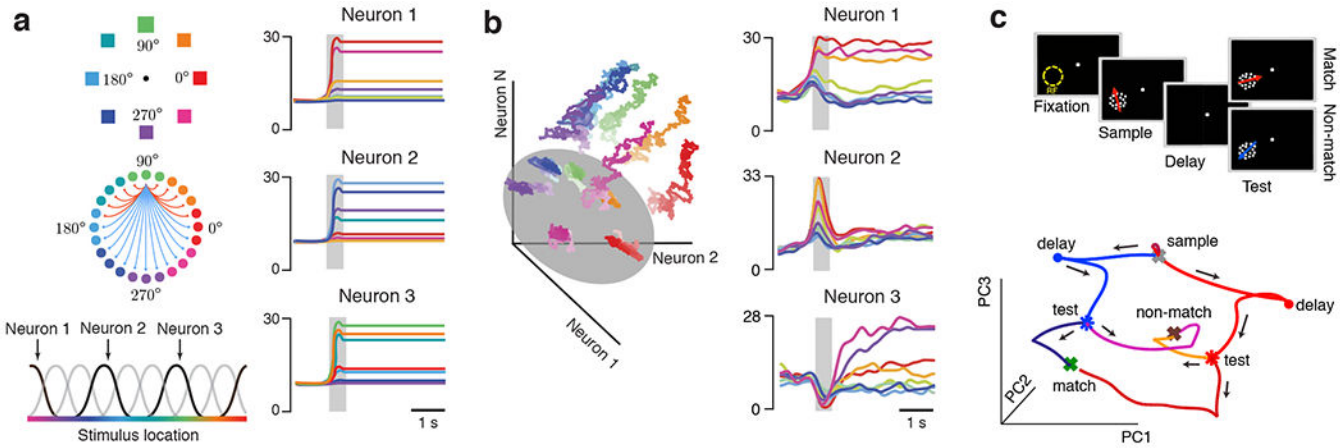


Figure 3. Low-dimensional task manifolds in heterogeneous responses of cortical neurons.

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 are colored 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[9,102,103]. A schematic representation of this activity in three model neurons in response to a stimulus in different locations (location is indicate by color) is shown in the right panels (y axis represents firing rate in Hz), 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 the same color 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, colored 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 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[109] (right panel). Colors 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 lateral intraparietal area and PFC as animals

complete the task[109]. Part b is adapted, with permission, from Ref.[29]. Part c is adapted, with permission from Ref.[109]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript