Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity

Mehrdad Jazayeri[1], Srdjan Ostojic[2]

Summary

The current exponential rise in recording capacity calls for new approaches for analysing and interpreting neural data. Effective dimensionality has emerged as a key concept for describing neural activity at the collective level, yet different studies rely on a variety of definitions of it. Here we focus on the complementary notions of intrinsic and embedding dimensionality, and argue that they provide a useful framework for extracting computational principles from data. Reviewing recent works, we propose that the intrinsic dimensionality reflects information about the latent variables encoded in collective activity, while embedding dimensionality reveals the manner in which this information is processed. Network models form an ideal substrate for testing more specifically the hypotheses on the computational principles reflected through intrinsic and embedding dimensionality.

Highlights

- Dimensionality has emerged as a key concept for describing collective neural activity
- Characterizing intrinsic and embedding dimensionality provides a guiding principle for interpreting underlying computations
- Intrinsic dimensionality reflects the nature of the information encoded in collective activity
- Embedding dimensionality describes how information is processed by neural circuits
- Network models provide a testing ground for the computational roles of intrinsic and embedding dimensionality
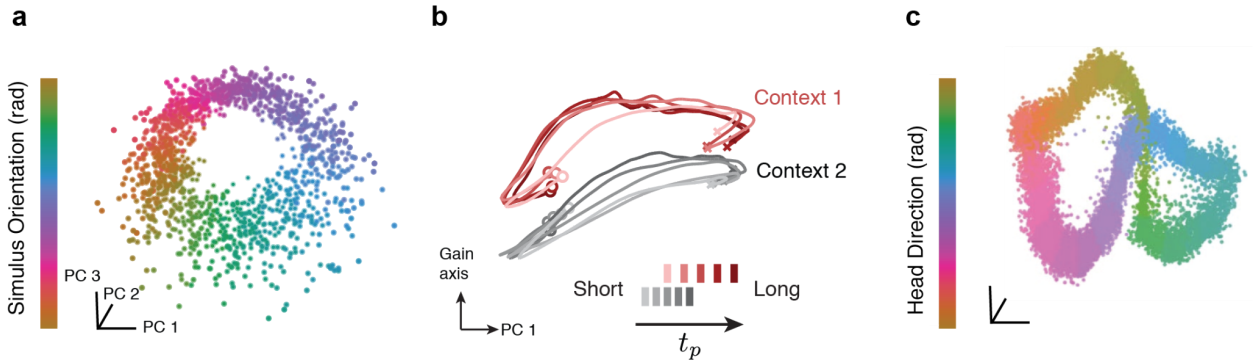
Addresses

[1]McGovern Institute for Brain Research, Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[2]Laboratoire de Neurosciences Cognitives, INSERM U960, École Normale Supérieure - PSL Research University, 75005 Paris, France

Introduction

Ongoing breakthroughs in tools and technologies for neural recording have unleashed an exponential rise in the number of neurons that can be simultaneously recorded in behaving animals [1]. The number of neurons we can record from in a brain area however, usually far exceeds the number of variables relevant to any given behavioral task. For example, in the primary motor cortex (M1), millions of neurons interact to control just a few muscles involved in a movement [2]. Similarly, interactions between large populations of neurons in the frontal and parietal cortex are thought to support simple decision-making tasks that involve just a few relevant variables [3–5]. This mismatch makes signals across neurons highly redundant and raises the question of how the population activity represents task-relevant variables [6–8].

To quantify this redundancy, and identify shared components of collective dynamics that reflect task variables, the effective dimensionality of population activity has emerged as a key concept. A central difficulty however is that different studies rely on different notions of dimensionality, so that a unifying picture is currently lacking. Here we focus on the distinction between *intrinsic* and *embedding* dimensionality of neural activity, and propose that it provides a guiding principle for interpreting neural computations. Reviewing recent experimental studies, we argue that intrinsic dimensionality reflects the nature of the information encoded in collective activity, while the embedding dimensionality describes the manner in which this information is processed by neural circuits. Systematically contrasting intrinsic and embedding dimensionality in different brain areas and behavioral settings is therefore a potentially powerful approach for generating hypotheses on neural computations. We finally discuss how the two types of dimensionality can be controlled within network models to implement specific computational hypotheses and produce further experimentally testable predictions.

Figure 1. Non-linear organization of neural activity. **a** Population activity in the mouse primary visual cortex in response to gratings of different orientations indicated in color. Projection on the first three principal components of data from [9]. **b** Population activity in the macaque dorsomedial frontal cortex reflecting flexible motor planning in a context-dependent timing task [10]. Despite the graded nature of the behavioral output across contexts (produced interval; t_p), the neural responses are organized non-linearly across contexts. **c** Population activity in the rat head-direction system [11]. Colors represent a computationally-inferred angular latent variable that maps directly onto head orientation. Panel **a** courtesy of NeuroMatch Academy [12] Pod 001 (Daniela Buchwald, Agustina Frechou, Habiba Noamany, Antonio Ortega)

Dimensionality of Neural Activity

The instantaneous firing rates across a sample of *N* neurons can be conceptualized as a point, or a *neural state*, in an *N*-dimensional Euclidean coordinate system known as the *state-space* where each axis represents the activity of one neuron. Within this framework, one can define the dimensionality in different ways. First is the *ambient dimensionality*, which is the dimension of the state space (*N*). Ambient dimensionsionality corresponds to the space of all possible neural states and thus does not reflect anything about the structure of neural activity.

Task-relevant computations usually depend on structured neural activity whose dimensionality is lower than that of the state space [13]. Quantifying this dimensionality is important as it bears information about the underlying variables and computations [11,13–16]. One common strategy to quantify this reduced dimensionality is to apply principal component analysis (PCA) to the neural data, from which one can extract patterns of activity that account for the largest amounts of variance. For instance, in a population of 100 neurons sampled during a specific task, a handful of principal components (PCs) may capture 80% of the total variance. This number of Euclidean dimensions that are required to capture the structured activity is referred to as the *embedding dimensionality* of the data.

While the embedding dimensionality carries information about structure, because of an assumption of linearity, it does not in general correspond to the number of independent degrees of freedom in the data (Fig. 1). For instance, a set of points organized along a ring has a single degree of freedom, but can be embedded in multiple dimensions (Box 1). More generally, the number of independent degrees of freedom is called the *intrinsic dimension*, and for non-linear structures it is distinct from the embedding dimensionality. Each independent degree of freedom is referred to as a *latent variable,* and for neural activity a key question is how latent variables relate to task variables.

Recent studies have combined simple linear methods such as PCA with more sophisticated nonlinear dimensionality reduction techniques to quantify both the intrinsic and embedding dimensionality of neural data providing insight into the nature of the underlying representations and computations [10,11,14,17–25]. For example, Chaudhuri et al. [11] analyzed neural activity associated with a population of head direction cells in the anterodorsal thalamic nucleus of mice during active foraging as well as rapid eye movement (REM) sleep. Using a novel nonlinear decoding strategy, they found that the population activity was organized around a ring structure with intrinsic dimensionality of 1, and that the position on the ring lawfully encoded animals' head direction. The ring, however, had a nontrivial topology, much like a twisted elastic band, so that the embedding dimensionality was much larger than one.

In another study, Low et al. [14] analyzed neural activity in the hippocampus and entorhinal cortex of mice during navigation. Using a novel nonlinear manifold inference technique, they were able to show that a large proportion of variance in the population activity was constrained to a manifold with intrinsic dimensionality of approximately 3, and that two of the dimensions corresponded to the animals' navigational space. However, due to its nonlinear embedding in the state space, the manifold was associated with a much larger embedding dimensionality. Intuitively, it was as if the 2D plane of navigation was twisted and stretched before being embedded in the neural state space.
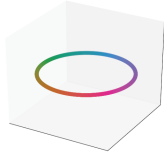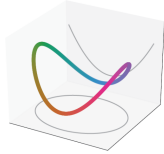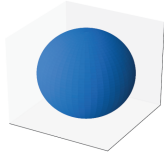
Together, these studies suggest that applying more systematically non-linear dimensionality reduction analyses to neural data may reveal the task-relevant latent variables more directly than through standard linear methods as PCA (Fig 1).

Box 1: Intrinsic vs embedding dimensionality

A premise of the analyses of dimensionality is that neural activity is restricted to one or several *manifolds* within the full state space. A manifold is a continuous subset of points characterized by two types of dimension: its intrinsic, or non-linear dimension, and its embedding, or linear dimension. The intrinsic dimension is the minimal number of continuous variables needed to parametrize the manifold. The embedding dimension is in contrast a measure of the number of dimensions explored by the manifold within the ambient Euclidean space.

A simple example of a non-linear manifold is a set of points forming a ring. The intrinsic dimension of a ring is one, as all the points can be parametrized by a single continuous variable, an angle along the ring. The embedding dimension of a ring however depends on how much it is twisted within the ambient space. A flat ring is embedded in two dimensions, while a warped ring is embedded in three or more dimensions, although it is still parametrized by a single angle. A spherical shell is a more complex example, where two parameters are needed to index every point, while the whole manifold is embedded in three or more dimensions.

For an idealized mathematical object such as a ring manifold, the concepts of linear and non-linear dimensionality are unambiguously defined. For real-world neural data, that is by nature noisy and incomplete, the definition and estimation of dimensionality instead hinges on the choice of a dimensionality reduction algorithm [22,26], leading to a vast heterogeneity across studies.

Intrinsic Dimension		Embedding Dimension
1		2
1		>2
2		3

Computational factors influencing intrinsic dimensionality

From a computational perspective, the intrinsic dimensionality of neural activity is largely determined by three sources of information: (1) incoming stimuli, (2) ongoing movements, and (3) the multitude of latent variables that characterize prior experiences and future expectations. This information, however, is not uniformly present in all brain areas. In early sensory areas, the intrinsic dimensionality is expected to be strongly associated with incoming stimuli. For example, responses to gratings in the primary visual cortex are organized along a circular variable corresponding to the grating orientation [9,18,21] (Fig 1a). Similarly, responses in the olfactory cortex reveal the latent organization of the chemical odour space [27]. The intrinsic dimensionality for more natural stimuli has not yet been systematically quantified, but it is likely that it would increase with the complexity of spatial [28] and/or temporal [29,30] dependencies.

On the other end of the spectrum are the neurons in the motor cortex and areas downstream that directly control muscle movements. Electrophysiology experiments and neural network modeling suggest that the dynamics of population activity in the motor cortex provide a potential basis set for

outgoing muscle-like commands [7,31–35]. Remarkably, signals in the primary motor cortex seem to carry information about the upcoming movement without regard to higher order latent variables such as context [36] or even the movement after the next [37]. As such, the intrinsic dimensionality in the late stages of the motor system seem to be strongly tied to ongoing movements (Fig 1b).

On the path from early sensory to late motor areas, neural signals go through a web of recurrent cortical and subcortical circuits that carry abstract information about sensory variables such as magnitude [38], navigation variables such as head direction (Fig 1c) [11] or position [20,23,25] , and other latent variables including behavioral context [4,10,15,39], contents of working memory [40], beliefs about rules [41], attentional state [42,43] and other internal states such as arousal and motivation [44]. Estimating the intrinsic dimensionality associated with these latent variables is a daunting task. First, long term dependencies in the structure of data can cause dimensionality to be exceedingly high [45]. Second, for cognitive tasks of any appreciable complexity, the latent variables the brain uses to control behavior are not known. One loosely articulated idea, motivated by humans' remarkable generalization capacity, is that these latent variables are organized such that knowledge about past experiences can be applied flexibly to new situations. One leading hypothesis from cognitive neuroscience and machine learning is that the brain decomposes experiences into explanatory factors such as objects, events, causal relationships, and other regularities that could support inference, and planning in new situations [46,47]. However, how these explanatory variables might map onto intrinsic variables and determine their dimensionality are among the most important unresolved questions at the intersection of machine learning and neuroscience [48]. Nonetheless, recent work has begun to make progress in this direction by developing sophisticated nonlinear dimensionality reduction techniques and applying them to large-scale neural data [10,11,14,19,49], and by using artificial neural network models to generate testable hypotheses for how neural systems might emulate complex cognitive behaviors [16,50,51].

Computational factors influencing embedding dimensionality

While the intrinsic dimension of the non-linear structure of population activity reflects stimuli, movements, and other core latent variables, its embedding into the Euclidean state space determines how these variables are processed and communicated. Different hypotheses about the nature of information processing in the brain can be casted as different forms of embedding. For example, in a system in which individual neurons and neural circuits play distinct functional roles, the state space can be parsimoniously partitioned into subspaces spanned by subsets of neurons. In contrast, when distinct sources of information are processed by partially overlapping groups of neurons (i.e., "mixed selectivity" [52]), the functional embeddings are no longer aligned with the principal axes of the state space.

Although the general principles that govern embedding dimensionality are not known, several computationally inspired hypotheses have been proposed. One dominant hypothesis is that the information in any given brain area is extracted by other areas and peripheral systems through a linear readout. A linear readout scheme is a weighted average of the activity across neurons and has an intuitive geometric interpretation: it is a projection of the neural activity in the full Euclidean state space along a specific direction. A sufficient (but not necessary) test of this hypothesis is to quantify the degree to which a linear decoder can extract desired information from population activity in a brain area.

An extension of this hypothesis is that embedding across a population of neurons is organized such that different linear decoders can extract information about different task-relevant variables without interference. For example, movement and movement preparation signals in monkeys' motor cortex reside in orthogonal subspaces [53]. This organization could ensure that preparation does not trigger movement [54,55], and may help protect previous motor memories [56]. Similar principles might govern interactions across cortical areas, by confining to orthogonal subspaces information that is private to an area and information that is communicated [57,58]. Orthogonal subspaces have also been implicated in the auditory cortex, both for reducing interference between new stimuli and old stimuli [40], and to enhance the representation of behaviorally relevant stimuli during task engagement [59]. Finally, orthogonal subspaces are thought to enable neural representations to disentangle working memory from motor preparation [60] and decision variables from confidence and contextual cues [4,10,24]. We note however, that despite strong enthusiasm, direct evidence that the brain processes and/or communicates information through orthogonal linear decoders is wanting.

In general, high-dimensional embeddings with more degrees of freedom can facilitate extraction of task-relevant variables without interference [52,61–63]. However, embedding information in arbitrarily high dimensional subspaces can have adverse effects for generalization [64]. To improve generalization, embeddings have to be appropriately constrained to capture the structural relationships and inherent invariances in the environment [10,38,65]. A theory for the ensuing trade-offs has been recently developed for the case where the activity is organized in multiple disjoint manifolds corresponding to different categories [66] and applied to interpret representations in the visual system [67] and in deep networks [68]. It is also possible to organize information embedding such that certain linear projection reflect information in more abstract form and therefore enable generalization, while others enable finer discrimination [15].

The idea of linear decodability, orthogonal subspaces, and their desiderata have straightforward interpretations for simple experiments in early sensory and late motor systems. For example, assuming that a core function of the early visual system is to categorize objects, modern models of vision are built to make object categories linearly decodable [69]. Similarly, in the motor system, open-loop movements such as ballistic reaches are ultimately constrained by how the activity is projected onto motor neurons, and models based on this idea can emulate neural responses in the motor cortex [33]. The utility of linear decodability however, becomes less clear for intermediate stages of information processing in higher brain areas that carry information about latent variables that support flexible mental computations. While an experimenter may apply linear decoders to find information about a hypothesized latent variable in a certain brain area, there is no a priori reason to assume that the brain relies on such decoders. In these situations, the computational constraint that likely determines the embedding properties of latent variables is their controllability [70]; i.e., the ability to flexibly use the underlying representation in the context of different policies, plans and action [71,72]. While there has been growing interest in the general notion of controllability in neural networks [73], our current conception of this problem is at its infancy [74] and more work is needed to understand how control may impact embedding dimensionality.

Network models as tools for testing hypotheses about intrinsic and embedding dimensionality

Large-scale recordings of neural activity together with innovative analysis techniques have paved the way toward more accurate measurements of intrinsic and embedding dimensionality in different brain areas and in a wide range of tasks. However, without concrete computational hypotheses, it could be extremely challenging to interpret measures of dimensionality. A powerful substrate for instantiating specific hypotheses and generating testable predictions are neural network models that can produce a desired behavior in a given task [51,75] or emulate patterns of neural activity associated with a certain brain area [76–78]. Comparison of neural network models to brain activity can narrow the space of hypotheses [50], and detailed analysis of candidate models can give insight into the underlying computational mechanisms [79]. For example, recent modeling efforts have found that the structure of neural activity in a wide range of brain areas is tightly linked to task requirements [3,10,39,80–84]. Systematic characterization of the underlying intrinsic and embedding dimensionality in network models however remains an open issue, and here we review promising paths forward.

Similarly to cortical activity, network models transform incoming sensory latent variables into behaviorally relevant outputs through a set of intermediate, abstract latent variables. A central question is how network dynamics expand the intrinsic dimensionality of activity to generate additional, internal latent variables that represent abstracted information not explicitly present in sensory stimuli. Recurrent dynamics are a key candidate mechanism for this process, as they can integrate past information to extract long-term dependencies relevant for the task. Recent works have examined mechanistically how the interplay between feed-forward and recurrent connectivity determines the dimensionality of activity [16,85], and shapes the dynamics of the abstract latent variables that emerge from recurrent dynamics [86]. For instance, ring-like manifolds, on which activity is represented by a single angular latent variable, can emerge from only weak structure in the connectivity [16,86], while more complex non-linear manifolds can be generated with additional structure [87]. Such modelling frameworks allow for a direct control of intrinsic and embedding dimensionality, and provide a rich testing ground for examining hypotheses on their computational roles.

A particularly important question pertains to the number and nature of intrinsic latent variables needed to implement a given computational task. More specifically, how should these latent variables interact, and be embedded into the collective activity to produce the desired input-output transform? A promising approach for addressing this question is to exploit model reduction techniques to reverse-engineer large recurrent networks trained on the task using techniques inspired by artificial intelligence [88]. Such model-reduction approaches extract key latent variables and represent their interactions in an interpretable manner. One example of model reduction is model distillation, which consists of retraining networks of minimal size based on the activity of the larger RNN [89]. This technique has been recently used to study a hierarchical inference task [90] currently tested on mice [91]. Schaeffer et al [90] applied model distillation to synthesize an interpretable reduced model that solved the task using two latent variables which captured the low-dimensional dynamics of the fully trained RNN. A second example of model reduction is mean-field analysis, that can in particular be used to extract the dynamics of latent variables directly from trained RNNs [92]. Remarkably, these latent dynamics can be represented in terms of effective circuits that describe how incoming sensory latent variables interact with internally generated latent variables [86]. Using the mean-field approach, large RNNs trained on various tasks can be therefore reduced to effective circuits of latent variables, opening the door to a systematic mapping from computational tasks onto latent dynamics [92].

Conclusion

In this review, we put forward the idea that intrinsic and embedding dimensionality provide potentially fruitful concepts for extracting computational principles from neural data. Testing this idea will require a systematic integration of non-linear analyses for dense neural recordings and computational models embodying computational and normative hypotheses, with the aim of determining the organization of latent variables underlying intrinsic dimensionality, and how they are put in use to implement computations.

References and recommended reading

1. Urai AE, Doiron B, Leifer AM, Churchland AK: **Large-scale neural recordings call for new insights to link brain and behavior**. *arXiv [q-bioNC]* 2021,

2. Russo AA, Bittner SR, Perkins SM, Seely JS, London BM, Lara AH, Miri A, Marshall NJ, Kohn A, Jessell TM, et al.: **Motor Cortex Embeds Muscle-like Commands in an Untangled Population Response**. *Neuron* 2018, **97**:953–966.e8.

3. Wang J, Narain D, Hosseini EA, Jazayeri M: **Flexible timing by temporal scaling of cortical responses**. *Nat Neurosci* 2018, **21**:102–110.

4. Mante V, Sussillo D, Shenoy KV, Newsome WT: **Context-dependent computation by recurrent dynamics in prefrontal cortex**. *Nature* 2013, **503**:78–84.

5. Hanks TD, Kopec CD, Brunton BW, Duan CA, Erlich JC, Brody CD: **Distinct relationships of parietal and prefrontal cortices to evidence accumulation**. *Nature* 2015, **520**:220–223.

6. Gao P, Ganguli S: **On simplicity and complexity in the brave new world of large-scale neuroscience**. *Curr Opin Neurobiol* 2015, **32**:148–155.

7. Gallego JA, Perich MG, Miller LE, Solla SA: **Neural Manifolds for the Control of Movement**. *Neuron* 2017, **94**:978–984.

8. Saxena S, Cunningham JP: **Towards the neural population doctrine**. *Curr Opin Neurobiol* 2019, **55**:103–111.

9.   Stringer C, Michaelos M, Tsyboulski D, Lindo SE, Pachitariu M: **High-precision coding in visual cortex**. *Cell* 2021, **184**:2767–2778.e15.

10. Remington ED, Narain D, Hosseini EA, Jazayeri M: **Flexible Sensorimotor Computations through Rapid Reconfiguration of Cortical Dynamics**. *Neuron* 2018, **98**:1005–1019.e5.

11. 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* 2019, **22**:1512–1520.

12. van Viegen T, Akrami A, Bonnen K, DeWitt E, Hyafil A, Ledmyr H, Lindsay GW, Mineault P, Murray JD, Pitkow X, et al.: **Neuromatch Academy: Teaching Computational Neuroscience with Global Accessibility**. *Trends Cogn Sci* 2021, **25**:535–538.

13. Gao P, Trautmann E, Yu B, Santhanam G, Ryu S, Shenoy K, Ganguli S: **A theory of multineuronal dimensionality, dynamics and measurement**. *bioRxiv* 2017, doi:10.1101/214262.

14. Low RJ, Lewallen S, Aronov D, Nevers R, Tank DW: **Probing variability in a cognitive map using manifold inference from neural dynamics**. *bioRxiv* 2018, doi:10.1101/418939.

15. Bernardi S, Benna MK, Rigotti M, Munuera J, Fusi S, Salzman CD: **The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex**. *Cell* 2020, **183**:954–967.e21.

16. Mastrogiuseppe F, Ostojic S: **Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks**. *Neuron* 2018, **99**:609–623.e29.

17. Stopfer M, Jayaraman V, Laurent G: **Intensity versus identity coding in an olfactory system**. *Neuron* 2003, **39**:991–1004.

18. Zhao Y, Park IM: **Variational Latent Gaussian Process for Recovering Single-Trial Dynamics from Population Spike Trains**. *Neural Comput* 2017, **29**:1293–1316.

19. Pandarinath C, O'Shea DJ, Collins J, Jozefowicz R, Stavisky SD, Kao JC, Trautmann EM, Kaufman MT, Ryu SI, Hochberg LR, et al.: **Inferring single-trial neural population dynamics using sequential auto-encoders**. *Nat Methods* 2018, **15**:805–815.

20. Rubin A, Sheintuch L, Brande-Eilat N, Pinchasof O, Rechavi Y, Geva N, Ziv Y: **Revealing neural correlates of behavior without behavioral measurements**. *Nat Commun* 2019, **10**:4745.

21. Ringach DL: **The geometry of masking in neural populations**. *Nat Commun* 2019, **10**:4879.

22. Altan E, Solla SA, Miller LE, Perreault EJ: **Estimating the dimensionality of the manifold underlying multi-electrode neural recordings**. *bioRxiv* 2020, doi:10.1101/2020.12.17.423196.

23. Gardner RJ, Hermansen E, Pachitariu M, Burak Y, Baas NA, Dunn BA, Moser M-B, Moser EI: **Toroidal topology of population activity in grid cells**. *bioRxiv* 2021, doi:10.1101/2021.02.25.432776.

24. Okazawa G, Hatch CE, Mancoo A, Machens CK: **The geometry of the representation of decision variable and stimulus difficulty in the parietal cortex**. *bioRxiv* [date unknown],

25. Nieh EH, Schottdorf M, Freeman NW, Low RJ, Lewallen S, Koay SA, Pinto L, Gauthier JL, Brody CD, Tank DW: **Geometry of abstract learned knowledge in the hippocampus**. *Nature* 2021,

26. Cunningham JP, Yu BM: **Dimensionality reduction for large-scale neural recordings**. *Nat Neurosci* 2014, **17**:1500–1509.

27. Pashkovski SL, Iurilli G, Brann D, Chicharro D, Drummey K, Franks KM, Panzeri S, Datta SR: **Structure and flexibility in cortical representations of odour space**. *Nature* 2020, **583**:253–258.

28. Pope P, Zhu C, Abdelkader A, Goldblum M, Goldstein T: **The Intrinsic Dimension of Images and Its Impact on Learning**. *arXiv [csCV]* 2021,

29. Goffinet J, Mooney R, Pearson J: **Inferring low-dimensional latent descriptions of animal vocalizations**. *bioRxiv* 2019,

30. Hénaff OJ, Goris RLT, Simoncelli EP: **Perceptual straightening of natural videos**. *Nat Neurosci* 2019, **22**:984–991.

31. Churchland MM, Cunningham JP, Kaufman MT, Foster JD, Nuyujukian P, Ryu SI, Shenoy KV: **Neural population dynamics during reaching**. *Nature* 2012, **487**:51–56.

32. Churchland MM, Cunningham JP, Kaufman MT, Ryu SI, Shenoy KV: **Cortical preparatory activity: representation of movement or first cog in a dynamical machine?** *Neuron* 2010, **68**:387–400.

33. Sussillo D, Churchland MM, Kaufman MT, Shenoy KV: **A neural network that finds a naturalistic solution for the production of muscle activity**. *Nat Neurosci* 2015, **18**:1025–1033.

34. Yan Y, Goodman JM, Moore DD, Solla SA, Bensmaia SJ: **Unexpected complexity of everyday manual behaviors**. *Nat Commun* 2020, **11**:3564.

35. Kato S, Kaplan HS, Schrödel T, Skora S, Lindsay TH, Yemini E, Lockery S, Zimmer M: **Global brain dynamics embed the motor command sequence of Caenorhabditis elegans**. *Cell* 2015, **163**:656–669.

36. Russo AA, Khajeh R, Bittner SR, Perkins SM, Cunningham JP, Abbott LF, Churchland MM: **Neural Trajectories in the Supplementary Motor Area and Motor Cortex Exhibit Distinct Geometries, Compatible with Different Classes of Computation**. *Neuron* 2020, **107**:745–758.e6.

37. Zimnik AJ, Churchland MM: **Independent generation of sequence elements by motor cortex**. *Nat Neurosci* 2021, doi:10.1038/s41593-021-00798-5.

38. Sheahan H, Luyckx F, Nelli S, Teupe C, Summerfield C: **Neural state space alignment for magnitude generalization in humans and recurrent networks**. *Neuron* 2021, doi:10.1016/j.neuron.2021.02.004.

39. Sohn H, Narain D, Meirhaeghe N, Jazayeri M: **Bayesian Computation through Cortical Latent Dynamics**. *Neuron* 2019, **103**:934–947.e5.

40. Libby A, Buschman TJ: **Rotational dynamics reduce interference between sensory and memory representations**. *Nat Neurosci* 2021, **24**:715–726.

41. Sarafyazd M, Jazayeri M: **Hierarchical reasoning by neural circuits in the frontal cortex**. *Science* 2019, **364**.

42. Ni AM, Ruff DA, Alberts JJ, Symmonds J, Cohen MR: **Learning and attention reveal a general relationship between population activity and behavior**. *Science* 2018, **359**:463–465.

43. Huang C, Ruff DA, Pyle R, Rosenbaum R, Cohen MR, Doiron B: **Circuit models of low-dimensional shared variability in cortical networks**. *Neuron* 2019, **101**:337–348.e4.

44. Cowley BR, Snyder AC, Acar K, Williamson RC, Yu BM, Smith MA: **Slow drift of neural activity as a signature of impulsivity in macaque visual and prefrontal cortex**. *Neuron* 2020, **108**:551–567.e8.

45. Bialek W: **What do we mean by the dimensionality of behavior?** *arXiv preprint arXiv:200809574* 2020,

46. Bengio Y, Courville A, Vincent P: **Representation learning: a review and new perspectives**. *IEEE Trans Pattern Anal Mach Intell* 2013, **35**:1798–1828.

47. Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ: **Building machines that learn and think like people**. *Behav Brain Sci* 2017, **40**.

48. Yang GR, Joglekar MR, Song HF, Newsome WT, Wang X-J: **Task representations in neural networks trained to perform many cognitive tasks**. *Nat Neurosci* 2019, **22**:297–306.

49. Meirhaeghe N, Sohn H, Jazayeri M: **A precise and adaptive neural mechanism for predictive temporal processing in the frontal cortex**. *bioRxiv* 2021,

50. Rajalingham R, Piccato A, Jazayeri M: **The role of mental simulation in primate physical inference abilities**. *bioRxiv* 2021,

51. Barak O: **Recurrent neural networks as versatile tools of neuroscience research**. *Curr Opin Neurobiol* 2017, **46**:1–6.

52. Rigotti M, Barak O, Warden MR, Wang X-J, Daw ND, Miller EK, Fusi S: **The importance of mixed selectivity in complex cognitive tasks**. *Nature* 2013, **497**:585–590.

53. Kaufman MT, Churchland MM, Ryu SI, Shenoy KV: **Cortical activity in the null space: permitting preparation without movement**. *Nat Neurosci* 2014, **17**:440–448.

54. Kao T-C, Sadabadi MS, Hennequin G: **Optimal anticipatory control as a theory of motor preparation: A thalamo-cortical circuit model**. *Neuron* 2021, **109**:1567–1581.e12.

55. Inagaki HK, Inagaki M, Romani S, Svoboda K: **Low-dimensional and monotonic preparatory activity in mouse anterior lateral motor cortex**. *J Neurosci* 2018,

38:4163–4185.

56. Sheahan HR, Franklin DW, Wolpert DM: **Motor Planning, Not Execution, Separates Motor Memories**. *Neuron* 2016, **92**:773–779.

57. Semedo JD, Zandvakili A, Machens CK, Yu BM, Kohn A: **Cortical Areas Interact through a Communication Subspace**. *Neuron* 2019, **102**:249–259.e4.

58. Srinath R, Ruff DA, Cohen MR: **Attention improves information flow between neuronal populations without changing the communication subspace**. *bioRxiv* 2021,

59. Bagur S, Averseng M, Elgueda D, David S, Fritz J, Yin P, Shamma S, Boubenec Y, Ostojic S: **Go/No-Go task engagement enhances population representation of target stimuli in primary auditory cortex**. *Nat Commun* 2018, **9**:2529.

60. Tang C, Herikstad R, Parthasarathy A, Libedinsky C, Yen S-C: **Minimally dependent activity subspaces for working memory and motor preparation in the lateral prefrontal cortex**. *Elife* 2020, **9**.

61. Cayco-Gajic NA, Clopath C, Silver RA: **Sparse synaptic connectivity is required for decorrelation and pattern separation in feedforward networks**. *Nat Commun* 2017, **8**:1116.

62. Cayco-Gajic NA, Silver RA: **Re-evaluating circuit mechanisms underlying pattern separation**. *Neuron* 2019, **101**:584–602.

63. Litwin-Kumar A, Harris KD, Axel R, Sompolinsky H, Abbott LF: **Optimal Degrees of Synaptic Connectivity**. *Neuron* 2017, **93**:1153–1164.e7.

64. Stringer C, Pachitariu M, Steinmetz N, Carandini M, Harris KD: **High-dimensional geometry of population responses in visual cortex**. *Nature* 2019, **571**:361–365.

65. DiCarlo JJ, Cox DD: **Untangling invariant object recognition**. *Trends Cogn Sci* 2007, **11**:333–341.

66. Chung SY, Lee DD, Sompolinsky H: **Classification and geometry of general perceptual manifolds**. *Physical Review X* 2018,

67. Froudarakis E, Cohen U, Diamantaki M, Walker EY: **Object manifold geometry across the mouse cortical visual hierarchy**. *bioRxiv* 2020,

68. Cohen U, Chung S, Lee DD, Sompolinsky H: **Separability and geometry of object manifolds in deep neural networks**. *Nat Commun* 2020, **11**:746.

69. Hong H, Yamins DLK, Majaj NJ, DiCarlo JJ: **Explicit information for category-orthogonal object properties increases along the ventral stream**. *Nat Neurosci* 2016, **19**:613–622.

70. Bengio E, Thomas V, Pineau J, Precup D, Bengio Y: **Independently Controllable Features**. *arXiv [csLG]* 2017,

71. Badre D, Bhandari A, Keglovits H, Kikumoto A: **The dimensionality of neural representations for control**. *Curr Opin Behav Sci* 2021, **38**:20–28.

72. Remington ED, Egger SW, Narain D, Wang J, Jazayeri M: **A Dynamical Systems Perspective on Flexible Motor Timing**. *Trends Cogn Sci* 2018, **22**:938–952.

73. Gu S, Pasqualetti F, Cieslak M, Telesford QK, Yu AB, Kahn AE, Medaglia JD, Vettel JM, Miller MB, Grafton ST, et al.: **Controllability of structural brain networks**. *Nat Commun* 2015, **6**:1–10.

74. Tu C, Rocha RP, Corbetta M, Zampieri S, Zorzi M, Suweis S: **Warnings and caveats in brain controllability**. *Neuroimage* 2018, **176**:83–91.

75. Sussillo D: **Neural circuits as computational dynamical systems**. *Curr Opin Neurobiol* 2014, **25**:156–163.

76. Rajan K, Harvey CD, Tank DW: **Recurrent Network Models of Sequence Generation and Memory**. *Neuron* 2016, **90**:128–142.

77. Pinto L, Rajan K, DePasquale B, Thiberge SY, Tank DW, Brody CD: **Task-Dependent Changes in the Large-Scale Dynamics and Necessity of Cortical Regions**. *Neuron* 2019, **104**:810–824.e9.

78. Andalman AS, Burns VM, Lovett-Barron M, Broxton M, Poole B, Yang SJ, Grosenick L, Lerner TN, Chen R, Benster T, et al.: **Neuronal Dynamics Regulating Brain and Behavioral State Transitions**. *Cell* 2019, **177**:970–985.e20.

79. Sussillo D, Barak O: **Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks**. *Neural Comput* 2013, **25**:626–649.

80. Schuessler F, Mastrogiuseppe F, Dubreuil A, Ostojic S, Barak O: **The interplay between randomness and structure during learning in RNNs**. *Adv Neural Inf Process Syst* 2020, **33**.

81. Farrell M, Recanatesi S, Moore T, Lajoie G: **Recurrent neural networks learn robust representations by dynamically balancing compression and expansion**. *bioRxiv* 2019,

82. Recanatesi S, Farrell M, Advani M, Moore T: **Dimensionality compression and expansion in Deep Neural Networks**. *arXiv preprint arXiv* 2019,

83. Recanatesi S, Farrell M, Lajoie G, Deneve S, Rigotti M, Shea-Brown E: **Predictive learning as a network mechanism for extracting low-dimensional latent space representations**. *Nat Commun* 2021, **12**:1417.

84. Masse NY, Yang GR, Song HF, Wang X-J, Freedman DJ: **Circuit mechanisms for the maintenance and manipulation of information in working memory**. *Nat Neurosci* 2019, **22**:1159–1167.

85. Schuessler F, Dubreuil A, Mastrogiuseppe F, Ostojic S, Barak O: **Dynamics of random recurrent networks with correlated low-rank structure**. *Physical Review Research* 2020, **2**:013111.

86. Beiran M, Dubreuil A, Valente A, Mastrogiuseppe F, Ostojic S: **Shaping dynamics with multiple populations in low-rank recurrent networks**. *arXiv [q-bioNC]* 2020,

87. Pollock E, Jazayeri M: **Engineering recurrent neural networks from task-relevant**

manifolds and dynamics. *PLoS Comput Biol* 2020, **16**:e1008128.

88. Yang GR, Wang X-J: **Artificial Neural Networks for Neuroscientists: A Primer**. *Neuron* 2020, **107**:1048–1070.

89. Hinton G, Vinyals O, Dean J: **Distilling the Knowledge in a Neural Network**. *arXiv [statML]* 2015,

90. Schaeffer R, Khona M, Meshulam L, International Brain Laboratory, Fiete IR: **Reverse-engineering Recurrent Neural Network solutions to a hierarchical inference task for mice**. *Cold Spring Harbor Laboratory* 2020, doi:10.1101/2020.06.09.142745.

91. International Brain Laboratory. Electronic address: churchland@cshl.edu, International Brain Laboratory: **An International Laboratory for Systems and Computational Neuroscience**. *Neuron* 2017, **96**:1213–1218.

92. Dubreuil AM, Valente A, Beiran M, Mastrogiuseppe F, Ostojic S: **Complementary roles of dimensionality and population structure in neural computations**. *biorxiv* 2020,