Signatures of task learning in neural representations

Harsha Gurnani[1] and N Alex Cayco Gajic[2]

[1] Department of Biology, University of Washington, Seattle, Washington, USA
[2] Laboratoire de Neuroscience Cognitives, Ecole Normale Supérieure, Université PSL, Paris, France

Corresponding author: Cayco Gajic, N Alex (natasha.cayco.gajic@ens.fr)

Highlights

- Task learning produces coordinated changes in neural population activity, both in neural circuits and in artificial neural networks (ANNs).
- Learning drives task-specific changes to the geometry of neural manifolds through expansion and compression of dimensionality and the orthogonalisation of behaviorally-relevant dimensions.
- New task computations are supported by changes to latent dynamics by modifying initial conditions of autonomous dynamics, realigning of slow and fast modes, or reshaping the attractor landscape.
- Studies in multi-task and continual learning in ANNs provides insight into how biological circuits could organize network activity to support multiple skills without catastrophic forgetting.

Abstract

While neural plasticity has long been studied as the basis of learning, the growth of large-scale neural recording techniques provides a unique opportunity to study how learning-induced activity changes are coordinated across neurons within the same circuit. These distributed changes can be understood through an evolution of the geometry of neural manifolds and latent dynamics underlying new computations. In parallel, studies of multi-task and continual learning in artificial neural networks hint at a tradeoff between non-interference and compositionality as guiding principles to understand how neural circuits flexibly support multiple behaviors. In this review, we highlight recent findings from both biological and artificial circuits that together form a new framework for understanding task learning at the population level.

Introduction

Learning to perform new tasks is fundamental to flexible and adaptive behavior in a changing environment. Neural correlates of learning have been observed as modifications of neural firing properties, including increases in single-neuron stimulus discriminability and the recruitment of new neurons encoding task-relevant information [1]–[5]. This view of learning mirrors classic principles of neural computation based on single neuron tuning properties. Yet the past decade of systems neuroscience has experienced a pivot towards more abstract population-level dynamics [6], [7], whose structure often reflects stimulus, task, and behavioral information, and can be used for computation [8]. In this review, we present recent experimental evidence of how population representations and dynamics evolve over learning as well as guiding theoretical principles from task-trained artificial neural networks (ANNs). We focus on three emerging themes: 1) learning as changes to geometric structure of neural manifolds, 2) learning as changes in the dynamics governing neural

population trajectories, and 3) learning multiple tasks, especially while avoiding catastrophic forgetting. Together these studies form the basis for a new conceptual framework which may provide a unifying picture of how task learning is expressed in population-level representations while admitting substantial degeneracy at the synaptic and neuronal levels.

From pairwise correlations to low-dimensional manifolds

A major theme in the study of neural populations is how interneuronal correlations shape population codes. Decades of theoretical work have shown that the information encoded in neural populations is limited when noise correlations (i.e., trial-to-trial covariability for a fixed stimulus) are aligned with signal correlations (i.e., correlations due to shared tuning properties) [9]. A natural question is therefore whether improvements in task performance over learning are caused by a change in correlation structure that increases the available stimulus information. In animals trained on associative learning tasks, an overall suppression of noise correlations for high signal correlation neuron pairs has indeed been observed for rewarded stimuli [2], [10]. Perceptual learning studies have reported a reduction of average noise correlations but the link to task performance remains unclear - either because the changes were non-specific [11] or because the subjects' choices were shown to be sub-optimal with respect to the population activity structure [12]. Importantly, recent work has suggested that noise correlations may change dynamically due to top-down feedback [13], and that knowledge of this modulation can be useful under certain decoding models [14].

Together, these mixed results point to the need to consider the global structure of variability in neural activity space, rather than average correlations between pairs of neurons [15]. Towards this end, a newer framework considers the low-dimensional structures in which population representations are embedded, often called neural "manifolds" (or when linear, also called "subspaces"). Neural manifold structure has been shown to constrain the speed of learning [16] and the manifold geometry underlying stable behaviors can remain stable over long timescales, even years [17]. While recent work using brain computer interfaces (BCIs) has argued that learning is constrained by pre-existing covariability structure [18], [19], one might expect the manifold structure to be flexible over long-timescale learning of new skills due to synaptic plasticity. Indeed, a follow up BCI study showed that animals were able to produce new neural patterns over several days of training (Figure 1A) [20]. In the following two sections, we focus on recent evidence from theoretical and experimental communities of how the geometry of neural manifolds could change to enable new task-specific computations.

Expansion and compression of dimensionality

A starting point to characterize changing manifold structure is to consider how the dimensionality of neural representations changes over learning. Both dimensionality expansion and compression have been reported at different computational stages, including at different layers of deep networks [21]–[23] and in recordings from different regions along the cortical hierarchy [24], [25]. This suggests that the expected relationship between changes in dimensionality and task performance depends on the chosen brain region. Experimental recordings in the motor cortex and associated regions have shown that motor skill learning is often accompanied by a reduction in the dimensionality of task-encoding activity subspace as well as in the variability of the embedded neural trajectories [26], [27]. While such low-dimensional representations may simply reflect task simplicity and reduced behavioral variability [28], there may also be normative pressures that favor the emergence of more compact representations. Intuitively, low-dimensional task representations allow for robust decoding in the presence of independent neuronal "noise", and can enable non-interfering representations of different skills to be stored in orthogonal low-dimensional subspaces.

While low-dimensional activity can be a signature of systems consolidation at the end of learning, high *initial* dimensionality could prove useful for probing new neural patterns in a search for optimal control [29]. In line with this idea, one study of BCI learning found an initially high contribution of “private” variability of individual motor cortical neurons to task performance, which reduced over trials as the population transitioned to using lower-dimensional “shared” variability for control [30]. By running either private or shared signals through the BCI decoder separately, Athalye and colleagues demonstrated that movements driven by the shared signal were faster and more direct, suggesting that the observed high-dimensional activity followed by compression reflected an implicit change in strategy for efficient goal-directed motor control. A compression of the dimensionality of stimulus representations was also observed in a recent classical conditioning study in the primate prefrontal cortex [31]. The authors argued that this compression could arise from the dual pressures of increased generalization and metabolic constraints [31], further showcasing the different tradeoffs inherent in dimensionality expansion and compression.

Additional intuition can be obtained from theoretical studies of learning in recurrent neural networks (RNNs), which are trained to minimize the error of a readout unit. For example, one recent study found that learning in RNNs trained on simple tasks induces weight changes that are low-rank [32], a form of connectivity structure which has previously been linked to low-dimensional neural activity [33]. But similar to [30], RNNs often show benefits of being initialized with high-dimensional activity in early epochs [32], [34]. Another key study showed that RNNs trained with stochastic gradient descent tend to compress representations specifically during the “decision” period, even in tasks where transient high-dimensional activity is computationally useful (Figure 1B) [35]. The authors pinpointed the cause of this dimensionality compression to the stochasticity inherent to the weight updates, which acts as an effective regularizer to enforce robust representations. While these studies show that low-dimensional neural representations could emerge as a result of task structure and gradient-based optimization, it remains an open question whether similar effects on dimensionality are found with biologically plausible plasticity rules which are local and may have task-irrelevant components [36]. Towards this end, one study has shown that dimensionality can be systematically affected by circuit motifs between pairs and triplets of neurons [37], suggesting a possible pathway for the regulation of global dimensionality through local synaptic plasticity.

Reorienting task-relevant dimensions

Beyond global changes in dimensionality, learning may drive specific changes in the geometry of neural representations. Examining how task-relevant dimensions reorient themselves over trials can provide insight regarding computational strategies that develop for solving the task. In particular, recent work has shown that neural population activity is often structured into orthogonal subspaces, enabling the circuit to separate sensory-driven from spontaneous activity patterns [38], to independently encode distinct task variables [39], [40], or to arrange dynamics in separate subspaces for context-specific computations [41]–[43]. Orthogonalisation of neural dimensions can also occur at different periods within a task in order to store non-interfering representations of past stimuli in short-term memory. This effect was recently observed in the auditory cortex during an implicit learning paradigm with auditory sequences: the optimal decoding axis for early stimuli evolved throughout sequence progression, and was eventually orthogonal to the original stimulus encoding dimensions [44]. This allowed previous stimuli in the sequence to be encoded by the same circuit within a “memory” subspace orthogonal to the “sensory” subspace. While Libby and Buschman report an increase in this orthogonalisation with passive exposure, the absence of behavioral readout precludes analysis of the timescale of statistical learning. Thus a more direct comparison of pre- and post-learning activity is needed to verify that such orthogonalisation is shaped by task requirements.

Rearranging activity into orthogonal subspaces can also be useful for selective information routing between brain regions. A single neuron can be viewed as a linear decoder, which projects upstream neural patterns onto a single dimension; any changes in upstream activity orthogonal to this dimension have no effect on its firing (Figure 1C). Similarly, a downstream population can be viewed as a linear subspace determined by the decoder axes of all its neurons. Neural circuits could thus regulate downstream responses by re-orienting activity from a “decoder-null” to a “decoder-potent” subspace to communicate the outcome of an internal computation to other brain regions [41], [42], [45], [46]. For example, learning of a visuomotor association task specifically orthogonalised responses for motor-associated stimuli in V1 [47]. Another study found that in animals trained on auditory Go/No-Go tasks, auditory cortical representations were reoriented during task engagement so that an inferred downstream stimulus decoder would observe enhanced activation during a target stimulus (which required a behavioral change), whereas reference stimuli would be indistinguishable from baseline activity [48]. In both studies, the orthogonalisation of task-relevant dimensions was not necessarily for an increase in stimulus discriminability. Rather, this selective and contextual re-orientation may correspond to an alignment with a decoder-potent subspace that could be used to trigger stronger downstream responses for behaviorally relevant stimuli (Figure 1C).

So far, we have focused on learning-induced changes to the geometry of neural manifolds or of decoder axes. An alternative perspective instead focuses on the low-dimensional subspaces that may control how neural trajectories unfold as a dynamical system [49], rather than low-dimensional subspaces in which neural activity is embedded. For example, rather than orthogonalization of dimensions encoding stimuli, a re-alignment of the dimensions dictating the flow of neural dynamics can be used to shape how upstream inputs are integrated by neural populations. Towards this end, [50] analyzed mouse V1 responses during a visual discrimination task, and inferred recurrent interactions and feedforward input by fitting autoregressive models to the data. They observed that task learning led to an increase in stimulus information in V1 without a concurrent increase in the inferred external input. Rather, improvements in V1 were better explained by enhanced temporal integration of task-relevant input after learning, and was achieved in the network by realigning the slowly decaying modes of recurrent dynamics with the most informative direction of network input (Figure 2A). This finding underscores the importance of not only considering learning-induced changes in manifold geometry, but also in the underlying dynamical processes that sculpt neural representations.

Reshaping the dynamics underlying task performance

Besides changes to the geometry of neural representations (e.g., realignment of stimulus-decoding dimensions; Figure 1C), learning can require changes to the dynamical structure that govern how population activity evolves under different conditions (e.g., realignment of slow and fast modes; Figure 2A). Activity-based methods focus on the structure of population activity in neural activity space, often with the implicit assumption that neural data points are drawn from some stationary distribution with no notion of time (e.g., PCA). On the other hand, dynamics-based methods consider the forces that determine how population activity changes from one time point to the next. This latter view takes insight from dynamical systems theory to ask how recurrent connectivity and neuron-intrinsic nonlinearities combine to shape changes in time-varying neural trajectories which implement population-level computations. These two views are not exclusive and can even be combined for a fuller view of task learning, as is demonstrated in recent work on motor learning: in the motor cortex, preparatory activity is often constrained to an “output-null” subspace orthogonal to movement-related activity [45], which is hypothesized to separate preparation from the execution of motor plans. However, orthogonal representations do not necessarily mean that the dynamics do not interact, as activity within the preparatory subspace can be causally important for the quality of succeeding movement kinematics [51], [52]. Refinement of motor skills may thus involve targeting preparatory activity to specific subspaces and setting appropriate “initial conditions” to trigger the relevant network dynamics during the movement phase (Figure 2B). This

hypothesis was tested by Perich and colleagues [53] who simultaneously recorded activity in primate dorsal premotor (PMd) and primary motor (M1) cortices during motor adaptation. By fitting predictive models of M1 neuron spiking using PMd and M1 population activity, the authors inferred that learning-induced changes in M1 activity were driven by upstream changes in planning-related activity in PMd. A more recent study observed structured rotations in preparatory activity states during curl force-field adaptation, which were specific to the direction of the curl field and the subset of targets showing behavioral signatures of adaptation [54]. These findings are consistent with a view that adapted neural dynamics in M1 can be generated by altered initial conditions without a change to the intrinsic M1 dynamical repertoire (although further experiments are needed to examine the potential contribution of small, correlated synaptic changes within M1 [55]).

In addition to setting new initial conditions for dynamics [53] or realignment of slow dynamical modes [50], learning could shape the internally generated attractor landscape. This was recently demonstrated in the anterolateral motor cortex (ALM) during a stimulus detection task [56]. Mice were trained to report the presence or absence of optogenetic stimulation in vibrissal somatosensory cortex (vS1) by licking left or right after a delay period, and to ignore further vS1 “distractor” stimulation, which occurred during the delay period. In distractor-trained mice, ALM network activity quickly recovered back to the correct choice-related activity. By contrast, in distractor-naive animals, even weak distractors led to more persistent perturbations of ALM activity, with frequent switches to the incorrect motor plan. Using task-trained ANNs, the authors inferred that robustness against distractors developed due to increased separation and stability of the two choice-related attractors, particularly in the late-delay period (Figure 2C). Together, these studies highlight how learning can shape task-relevant dynamics via rearrangement of input and recurrent interactions.

Learning multiple tasks via non-interference and compositionality

Beyond learning of individual tasks, a growing body of work asks how neural circuits can flexibly support multiple simultaneously learned tasks. Expanding from the ideas above, one effective strategy to minimize interference between tasks (or “contexts”) is to use orthogonal task representations. This can take the form of separating task features along different dimensions for selective integration [39]; compression of stimulus representations along task-irrelevant dimensions [57]; or at its extreme, the use of non-overlapping neural populations in different contexts [58], [59]. Such a strategy is most useful when those tasks don’t share common computations and/or require knowledge of independent stimulus features. But as organisms acquire a vast, flexible repertoire of overlapping skills over their lifetimes, it becomes more efficient to instead decompose tasks into modular operations that can be recombined in many ways [60]. Such compositionality has been observed in ANNs trained simultaneously on 15-20 tasks, both in the algebraic relationships of different task representations [61] as well as regarding the reuse of dynamical structure for similar tasks [62], [63]. While these results were observed without architectural constraints, a bias for compositional representations could be imposed through structural bottlenecks and gating inputs to force shared representations [64].

A related perspective emerges in recent work that focuses on sequential (as opposed to simultaneous) learning: how does new task learning interfere with previously acquired skills? In particular, ANNs are notoriously susceptible to “catastrophic forgetting” of prior tasks when new tasks are learned. Multiple solutions to catastrophic forgetting have recently been proposed under the umbrella term of continual learning [65], [66], including penalizing weight changes that would incur a loss in performance on previous tasks [67], orthogonalizing weight updates to encourage non-interfering representations (Figure 3A) [68]–[71] and contextual gating induced by Hebbian learning rules [72]. While these theory-driven hypotheses remain challenging to validate in neural data, one recent study observed an interesting parallel when monkeys were

sequentially trained on two BCI mappings A and B [73]. After learning the second map B, neural activity during re-exposure to map A had moved along the “null space” of task A (i.e., orthogonal to map A) and didn’t impede their performance. However these shifted neural activity patterns could also produce improved control signals via map B and could serve as a “memory trace” for B (Figure 3B). More broadly, sequential learning studies in primates have observed that information about previous training is reflected in the geometry of task-specific representations [54], [74], echoing findings that the emergent solutions found by ANNs generally depend on pre-training (and can even reorganize representations corresponding to previously learnt tasks without a drop in performance) [62], [75]. However, a key difference between ANNs and biological brains is the high degree of (hierarchical) modularity found in the latter, with disparate learning rules identified across regions, suggesting that a distinct strategy may be at play to enable learning throughout an organism’s lifetime. More work needs to be done, both in artificial and biological neural networks, to fully understand the tradeoffs between non-interference and compositionality as guiding principles for learning multiple tasks.

Outlook and challenges ahead

Viewing task learning as shaping of population dynamics to support new computations provides a powerful framework to interpret the diverse learning-induced changes that have been observed in neural data. While we mainly focused on neocortical circuits in this perspective article, these principles extend to learning-related reorganization in other brain regions such as hippocampal [76], cerebellar [26], [77] and limbic [78] circuits, and need to be bridged with decades-long insights on synaptic and intrinsic plasticity mechanisms that enable learning [79], [80]. To forge ahead, several experimental and theoretical challenges need to be overcome. Longitudinal monitoring of the same neural population throughout learning is critical for teasing apart refinement of existing motifs from emergence of new features of population activity [81]–[83]. Data-analysis tools such as alignment [17], [84], [85] or identification of latent structure across sessions [86], [87], and disentangling input from recurrent contributions [88], [89] will be crucial to track distributed changes in neural representations over learning. Reporting training history and inferring the implicit strategies of animals will help link these neural changes to the behaviors they support, and account for inter-individual variability [74], [90], [91]. Lastly, leveraging further insights from learning in ANNs requires a better understanding of how different (or similar) gradient-based optimization is to biological learning [19], [92] as well as an examination of the dependence of learnt task representations on specific learning rules [93], [94]. Together, these advances will provide a better understanding of the mechanisms underlying behavioral flexibility, guiding future work on lifelong learning in biological and artificial agents.

Acknowledgements

We thank Angus Chadwick, Ashok Litwin-Kumar, Francesca Mastrogiuseppe, Arthur Pellegrino, and Heike Stein for feedback on the manuscript. This work was supported by a Schmidt Science Fellowship awarded to H.G. and the Agence Nationale de la Recherche (N.A.C.G., ANR-17-EURE-0017).

Declaration of Interest

The authors have declared no competing interests.

Bibliography

[1] J. Poort *et al.*, “Learning Enhances Sensory and Multiple Non-sensory Representations in Primary Visual Cortex,” *Neuron*, vol. 86, no. 6, pp. 1478–1490, Jun. 2015, doi: 10.1016/j.neuron.2015.05.037.

[2] P. M. Goltstein, G. T. Meijer, and C. M. Pennartz, “Conditioning sharpens the spatial representation of rewarded stimuli in mouse primary visual cortex,” *eLife*, vol. 7, p. e37683, Sep. 2018, doi: 10.7554/eLife.37683.

[3] S. Reinert, M. Hübener, T. Bonhoeffer, and P. M. Goltstein, “Mouse prefrontal cortex represents learned rules for categorization,” *Nature*, vol. 593, no. 7859, Art. no. 7859, May 2021, doi: 10.1038/s41586-021-03452-z.

[4] J. W. Schumacher, M. K. McCann, K. J. Maximov, and D. Fitzpatrick, “Selective enhancement of neural coding in V1 underlies fine-discrimination learning in tree shrew,” *Curr. Biol.*, vol. 32, no. 15, pp. 3245-3260.e5, Aug. 2022, doi: 10.1016/j.cub.2022.06.009.

[5] J. U. Henschke *et al.*, “Reward Association Enhances Stimulus-Specific Representations in Primary Visual Cortex,” *Curr. Biol.*, vol. 30, no. 10, pp. 1866-1880.e5, May 2020, doi: 10.1016/j.cub.2020.03.018.

[6] D. L. Barack and J. W. Krakauer, “Two views on the cognitive brain,” *Nat. Rev. Neurosci.*, vol. 22, no. 6, Art. no. 6, Jun. 2021, doi: 10.1038/s41583-021-00448-6.

[7] S. Saxena and J. P. Cunningham, “Towards the neural population doctrine,” *Curr. Opin. Neurobiol.*, vol. 55, pp. 103–111, Apr. 2019, doi: 10.1016/j.conb.2019.02.002.

[8] S. Vyas, M. D. Golub, D. Sussillo, and K. V. Shenoy, “Computation Through Neural Population Dynamics,” *Annu. Rev. Neurosci.*, vol. 43, no. 1, pp. 249–275, 2020, doi: 10.1146/annurev-neuro-092619-094115.

[9] A. Kohn, R. Coen-Cagli, I. Kanitscheider, and A. Pouget, “Correlations and Neuronal Population Information,” *Annu. Rev. Neurosci.*, vol. 39, no. 1, pp. 237–256, 2016, doi: 10.1146/annurev-neuro-070815-013851.

[10] J. M. Jeanne, T. O. Sharpee, and T. Q. Gentner, “Associative Learning Enhances Population Coding by Inverting Interneuronal Correlation Patterns,” *Neuron*, vol. 78, no. 2, pp. 352–363, Apr. 2013, doi: 10.1016/j.neuron.2013.02.023.

[11] Y. Gu *et al.*, “Perceptual Learning Reduces Interneuronal Correlations in Macaque Visual Cortex,” *Neuron*, vol. 71, no. 4, pp. 750–761, Aug. 2011, doi: 10.1016/j.neuron.2011.06.015.

[12] A. M. Ni, D. A. Ruff, J. J. Alberts, J. Symmonds, and M. R. Cohen, “Learning and attention reveal a general relationship between population activity and behavior,” *Science*, vol. 359, no. 6374, pp. 463–465, Jan. 2018, doi: 10.1126/science.aao0284.

[13] A. G. Bondy, R. M. Haefner, and B. G. Cumming, “Feedback determines the structure of correlated variability in primary visual cortex,” *Nat. Neurosci.*, vol. 21, no. 4, Art. no. 4, Apr. 2018, doi: 10.1038/s41593-018-0089-1.

[14] C. Haimerl, D. A. Ruff, M. R. Cohen, C. Savin, and E. P. Simoncelli, “Targeted comodulation supports flexible and accurate decoding in V1.” bioRxiv, p. 2021.02.23.432351, Feb. 23, 2021. doi: 10.1101/2021.02.23.432351.

[15] N. Kriegeskorte and X.-X. Wei, “Neural tuning and representational geometry,” *Nat. Rev. Neurosci.*, vol. 22, no. 11, Art. no. 11, Nov. 2021, doi: 10.1038/s41583-021-00502-3.

[16] P. T. Sadtler *et al.*, “Neural constraints on learning,” *Nature*, vol. 512, no. 7515, Art. no. 7515, Aug. 2014, doi: 10.1038/nature13665.

[17] J. A. Gallego, M. G. Perich, R. H. Chowdhury, S. A. Solla, and L. E. Miller, “Long-term stability of cortical population dynamics underlying consistent behavior,” *Nat. Neurosci.*, vol. 23, no. 2, Art. no. 2, Feb. 2020, doi: 10.1038/s41593-019-0555-4.

[18] M. D. Golub *et al.*, “Learning by neural reassociation,” *Nat. Neurosci.*, vol. 21, no. 4, Art. no. 4, Apr. 2018, doi: 10.1038/s41593-018-0095-3.

[19] J. A. Hennig, E. R. Oby, D. M. Losey, A. P. Batista, B. M. Yu, and S. M. Chase, “How learning unfolds in the brain: toward an optimization view,” *Neuron*, vol. 109, no. 23, pp. 3720–3735, Dec. 2021, doi: 10.1016/j.neuron.2021.09.005.

[20] E. R. Oby *et al.*, “New neural activity patterns emerge with long-term learning,” *Proc. Natl. Acad. Sci.*, vol. 116, no. 30, pp. 15210–15215, Jul. 2019, doi: 10.1073/pnas.1820296116.

[21] A. Ansuini, A. Laio, J. H. Macke, and D. Zoccolan, “Intrinsic dimension of data representations in deep

neural networks," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2019. Accessed: Apr. 14, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/hash/cfcce0621b49c983991ead4c3d4d3b6b-Abstract.html
[22] S. Recanatesi, M. Farrell, M. Advani, T. Moore, G. Lajoie, and E. Shea-Brown, "Dimensionality compression and expansion in Deep Neural Networks." arXiv, Oct. 27, 2019. doi: 10.48550/arXiv.1906.00443.
[23] U. Cohen, S. Chung, D. D. Lee, and H. Sompolinsky, "Separability and geometry of object manifolds in deep neural networks," *Nat. Commun.*, vol. 11, no. 1, Art. no. 1, Feb. 2020, doi: 10.1038/s41467-020-14578-5.
[24] E. Froudarakis *et al.*, "Object manifold geometry across the mouse cortical visual hierarchy." bioRxiv, p. 2020.08.20.258798, Sep. 22, 2021. doi: 10.1101/2020.08.20.258798.
[25] S. L. Brincat, M. Siegel, C. von Nicolai, and E. K. Miller, "Gradual progression from sensory to task-related processing in cerebral cortex," *Proc. Natl. Acad. Sci.*, vol. 115, no. 30, pp. E7202–E7211, Jul. 2018, doi: 10.1073/pnas.1717075115.
[26] M. J. Wagner *et al.*, "Shared Cortex-Cerebellum Dynamics in the Execution and Learning of a Motor Task," *Cell*, vol. 177, no. 3, pp. 669-682.e24, Apr. 2019, doi: 10.1016/j.cell.2019.02.019.
[27] J. Kim, A. Joshi, L. Frank, and K. Ganguly, "Cortical–hippocampal coupling during manifold exploration in motor cortex," *Nature*, vol. 613, no. 7942, Art. no. 7942, Jan. 2023, doi: 10.1038/s41586-022-05533-z.
[28] P. Gao *et al.*, "A theory of multineuronal dimensionality, dynamics and measurement." bioRxiv, p. 214262, Nov. 12, 2017. doi: 10.1101/214262.
[29] A. K. Dhawale, M. A. Smith, and B. P. Ölveczky, "The Role of Variability in Motor Learning," *Annu. Rev. Neurosci.*, vol. 40, no. 1, pp. 479–498, 2017, doi: 10.1146/annurev-neuro-072116-031548.
[30] V. R. Athalye, K. Ganguly, R. M. Costa, and J. M. Carmena, "Emergence of Coordinated Neural Dynamics Underlies Neuroprosthetic Learning and Skillful Control," *Neuron*, vol. 93, no. 4, pp. 955-970.e5, Feb. 2017, doi: 10.1016/j.neuron.2017.01.016.
[31] M. J. Wójcik *et al.*, "Learning shapes neural geometry in the prefrontal cortex." bioRxiv, p. 2023.04.24.538054, Apr. 24, 2023. doi: 10.1101/2023.04.24.538054.
[32] F. Schuessler, F. Mastrogiuseppe, A. Dubreuil, S. Ostojic, and O. Barak, "The interplay between randomness and structure during learning in RNNs." arXiv, May 13, 2021. doi: 10.48550/arXiv.2006.11036.
[33] F. Mastrogiuseppe and S. Ostojic, "Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks," *Neuron*, vol. 99, no. 3, pp. 609-623.e29, Aug. 2018, doi: 10.1016/j.neuron.2018.07.003.
[34] D. Sussillo and L. F. Abbott, "Generating Coherent Patterns of Activity from Chaotic Neural Networks," *Neuron*, vol. 63, no. 4, pp. 544–557, Aug. 2009, doi: 10.1016/j.neuron.2009.07.018.
[35] M. Farrell, S. Recanatesi, T. Moore, G. Lajoie, and E. Shea-Brown, "Gradient-based learning drives robust representations in recurrent neural networks by balancing compression and expansion," *Nat. Mach. Intell.*, vol. 4, no. 6, Art. no. 6, Jun. 2022, doi: 10.1038/s42256-022-00498-0.
[36] D. V. Raman, A. P. Rotondo, and T. O'Leary, "Fundamental bounds on learning performance in neural circuits," *Proc. Natl. Acad. Sci.*, vol. 116, no. 21, pp. 10537–10546, May 2019, doi: 10.1073/pnas.1813416116.
[37] S. Recanatesi, G. K. Ocker, M. A. Buice, and E. Shea-Brown, "Dimensionality in recurrent spiking networks: Global trends in activity and local origins in connectivity," *PLOS Comput. Biol.*, vol. 15, no. 7, p. e1006446, Jul. 2019, doi: 10.1371/journal.pcbi.1006446.
[38] L. Avitan and C. Stringer, "Not so spontaneous: Multi-dimensional representations of behaviors and context in sensory areas," *Neuron*, vol. 110, no. 19, pp. 3064–3075, Oct. 2022, doi: 10.1016/j.neuron.2022.06.019.
[39] V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome, "Context-dependent computation by recurrent dynamics in prefrontal cortex," *Nature*, vol. 503, no. 7474, Art. no. 7474, Nov. 2013, doi: 10.1038/nature12742.
[40] Y. Xie *et al.*, "Geometry of sequence working memory in macaque prefrontal cortex," *Science*, vol. 375, no. 6581, pp. 632–639, Feb. 2022, doi: 10.1126/science.abm0204.
[41] S. B. M. Yoo and B. Y. Hayden, "The Transition from Evaluation to Selection Involves Neural Subspace

Reorganization in Core Reward Regions," *Neuron*, vol. 105, no. 4, pp. 712-724.e4, Feb. 2020, doi: 10.1016/j.neuron.2019.11.013.
[42] M. T. Kaufman, M. M. Churchland, S. I. Ryu, and K. V. Shenoy, "Cortical activity in the null space: permitting preparation without movement," *Nat. Neurosci.*, vol. 17, no. 3, Art. no. 3, Mar. 2014, doi: 10.1038/nn.3643.
[43] F. Lanore, N. A. Cayco-Gajic, H. Gurnani, D. Coyle, and R. A. Silver, "Cerebellar granule cell axons support high-dimensional representations," *Nat. Neurosci.*, vol. 24, no. 8, pp. 1142–1150, Aug. 2021, doi: 10.1038/s41593-021-00873-x.
[44] A. Libby and T. J. Buschman, "Rotational dynamics reduce interference between sensory and memory representations," *Nat. Neurosci.*, vol. 24, no. 5, Art. no. 5, May 2021, doi: 10.1038/s41593-021-00821-9.
[45] G. F. Elsayed, A. H. Lara, M. T. Kaufman, M. M. Churchland, and J. P. Cunningham, "Reorganization between preparatory and movement population responses in motor cortex," *Nat. Commun.*, vol. 7, no. 1, Art. no. 1, Oct. 2016, doi: 10.1038/ncomms13239.
[46] J. D. Semedo, A. Zandvakili, C. K. Machens, B. M. Yu, and A. Kohn, "Cortical Areas Interact through a Communication Subspace," *Neuron*, vol. 102, no. 1, pp. 249-259.e4, Apr. 2019, doi: 10.1016/j.neuron.2019.01.026.
[47] S. W. Failor, M. Carandini, and K. D. Harris, "Visuomotor association orthogonalizes visual cortical population codes." bioRxiv, p. 2021.05.23.445338, Nov. 28, 2022. doi: 10.1101/2021.05.23.445338.
[48] S. Bagur *et al.*, "Go/No-Go task engagement enhances population representation of target stimuli in primary auditory cortex," *Nat. Commun.*, vol. 9, no. 1, Art. no. 1, Jun. 2018, doi: 10.1038/s41467-018-04839-9.
[49] L. Duncker and M. Sahani, "Dynamics on the manifold: Identifying computational dynamical activity from neural population recordings," *Curr. Opin. Neurobiol.*, vol. 70, pp. 163–170, Oct. 2021, doi: 10.1016/j.conb.2021.10.014.
[50] A. Chadwick *et al.*, "Learning shapes cortical dynamics to enhance integration of relevant sensory input," *Neuron*, vol. 111, no. 1, pp. 106-120.e10, Jan. 2023, doi: 10.1016/j.neuron.2022.10.001.
[51] A. Afshar, G. Santhanam, B. M. Yu, S. I. Ryu, M. Sahani, and K. V. Shenoy, "Single-Trial Neural Correlates of Arm Movement Preparation," *Neuron*, vol. 71, no. 3, pp. 555–564, Aug. 2011, doi: 10.1016/j.neuron.2011.05.047.
[52] K. V. Shenoy, M. Sahani, and M. M. Churchland, "Cortical Control of Arm Movements: A Dynamical Systems Perspective," *Annu. Rev. Neurosci.*, vol. 36, no. 1, pp. 337–359, 2013, doi: 10.1146/annurev-neuro-062111-150509.
[53] M. G. Perich, J. A. Gallego, and L. E. Miller, "A Neural Population Mechanism for Rapid Learning," *Neuron*, vol. 100, no. 4, pp. 964-976.e7, Nov. 2018, doi: 10.1016/j.neuron.2018.09.030.
[54] X. Sun *et al.*, "Cortical preparatory activity indexes learned motor memories," *Nature*, vol. 602, no. 7896, Art. no. 7896, Feb. 2022, doi: 10.1038/s41586-021-04329-x.
[55] B. Feulner, M. G. Perich, R. H. Chowdhury, L. E. Miller, J. A. Gallego, and C. Clopath, "Small, correlated changes in synaptic connectivity may facilitate rapid motor learning," *Nat. Commun.*, vol. 13, no. 1, Art. no. 1, Sep. 2022, doi: 10.1038/s41467-022-32646-w.
[56] A. Finkelstein, L. Fontolan, M. N. Economo, N. Li, S. Romani, and K. Svoboda, "Attractor dynamics gate cortical information flow during decision-making," *Nat. Neurosci.*, vol. 24, no. 6, Art. no. 6, Jun. 2021, doi: 10.1038/s41593-021-00840-6.
[57] T. Flesch, K. Juechems, T. Dumbalska, A. Saxe, and C. Summerfield, "Orthogonal representations for robust context-dependent task performance in brains and neural networks," *Neuron*, vol. 110, no. 7, pp. 1258-1270.e11, Apr. 2022, doi: 10.1016/j.neuron.2022.01.005.
[58] J. J. Lee, M. Krumin, K. D. Harris, and M. Carandini, "Task specificity in mouse parietal cortex," *Neuron*, vol. 110, no. 18, pp. 2961-2969.e5, Sep. 2022, doi: 10.1016/j.neuron.2022.07.017.
[59] J. E. Roy, M. Riesenhuber, T. Poggio, and E. K. Miller, "Prefrontal Cortex Activity during Flexible Categorization," *J. Neurosci.*, vol. 30, no. 25, pp. 8519–8528, Jun. 2010, doi: 10.1523/JNEUROSCI.4837-09.2010.
[60] G. R. Yang, M. W. Cole, and K. Rajan, "How to study the neural mechanisms of multiple tasks," *Curr. Opin. Behav. Sci.*, vol. 29, pp. 134–143, Oct. 2019, doi: 10.1016/j.cobeha.2019.07.001.
[61] G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome, and X.-J. Wang, "Task representations in neural networks trained to perform many cognitive tasks," *Nat. Neurosci.*, vol. 22, no. 2, Art. no. 2, Feb. 2019,

doi: 10.1038/s41593-018-0310-2.
[62] L. Driscoll, K. Shenoy, and D. Sussillo, “Flexible multitask computation in recurrent networks utilizes shared dynamical motifs.” bioRxiv, p. 2022.08.15.503870, Aug. 15, 2022. doi: 10.1101/2022.08.15.503870.
[63] V. Goudar, B. Peysakhovich, D. J. Freedman, E. A. Buffalo, and X.-J. Wang, “Schema formation in a neural population subspace underlies learning-to-learn in flexible sensorimotor problem-solving,” *Nat. Neurosci.*, vol. 26, no. 5, Art. no. 5, May 2023, doi: 10.1038/s41593-023-01293-9.
[64] A. M. Saxe, S. Sodhani, and S. Lewallen, “The Neural Race Reduction: Dynamics of Abstraction in Gated Networks.” arXiv, Jul. 21, 2022. doi: 10.48550/arXiv.2207.10430.
[65] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, “Continual Lifelong Learning with Neural Networks: A Review,” *Neural Netw.*, vol. 113, pp. 54–71, May 2019, doi: 10.1016/j.neunet.2019.01.012.
[66] T. Flesch, A. Saxe, and C. Summerfield, “Continual task learning in natural and artificial agents,” *Trends Neurosci.*, vol. 46, no. 3, pp. 199–210, Mar. 2023, doi: 10.1016/j.tins.2022.12.006.
[67] J. Kirkpatrick *et al.*, “Overcoming catastrophic forgetting in neural networks,” *Proc. Natl. Acad. Sci.*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017, doi: 10.1073/pnas.1611835114.
[68] L. Duncker, L. Driscoll, K. V. Shenoy, M. Sahani, and D. Sussillo, “Organizing recurrent network dynamics by task-computation to enable continual learning,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 14387–14397. Accessed: Dec. 15, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/a576eafbce762079f7d1f77fca1c5cc2-Abstract.html
[69] A. Chaudhry, N. Khan, P. Dokania, and P. Torr, “Continual Learning in Low-rank Orthogonal Subspaces,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 9900–9911. Accessed: Apr. 14, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/hash/70d85f35a1fdc0ab701ff78779306407-Abstract.html
[70] G. Zeng, Y. Chen, B. Cui, and S. Yu, “Continual learning of context-dependent processing in neural networks,” *Nat. Mach. Intell.*, vol. 1, no. 8, Art. no. 8, Aug. 2019, doi: 10.1038/s42256-019-0080-x.
[71] M. Farajtabar, N. Azizan, A. Mott, and A. Li, “Orthogonal Gradient Descent for Continual Learning,” in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, PMLR, Jun. 2020, pp. 3762–3773. Accessed: Mar. 21, 2023. [Online]. Available: https://proceedings.mlr.press/v108/farajtabar20a.html
[72] T. Flesch, D. G. Nagy, A. Saxe, and C. Summerfield, “Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals,” *PLOS Comput. Biol.*, vol. 19, no. 1, p. e1010808, Jan. 2023, doi: 10.1371/journal.pcbi.1010808.
[73] D. M. Losey *et al.*, “Learning alters neural activity to simultaneously support memory and action.” bioRxiv, p. 2022.07.05.498856, Jul. 06, 2022. doi: 10.1101/2022.07.05.498856.
[74] K. W. Latimer and D. J. Freedman, “Low-dimensional encoding of decisions in parietal cortex reflects long-term training history,” *Nat. Commun.*, vol. 14, no. 1, Art. no. 1, Feb. 2023, doi: 10.1038/s41467-023-36554-5.
[75] D. R. Kepple, R. Engelken, and K. Rajan, “Curriculum learning as a tool to uncover learning principles in the brain,” in *International Conference on Learning Representations*, Jan. 2022. Accessed: Mar. 21, 2023. [Online]. Available: https://openreview.net/forum?id=TpJMvo0_pu-
[76] V. Samborska, J. L. Butler, M. E. Walton, T. E. J. Behrens, and T. Akam, “Complementary task representations in hippocampus and prefrontal cortex for generalizing the structure of problems,” *Nat. Neurosci.*, vol. 25, no. 10, Art. no. 10, Oct. 2022, doi: 10.1038/s41593-022-01149-8.
[77] S. Muscinelli, M. Wagner, and A. Litwin-Kumar, “Optimal routing to cerebellum-like structures,” Neuroscience, preprint, Feb. 2022. doi: 10.1101/2022.02.10.480014.
[78] B. F. Grewe *et al.*, “Neural ensemble dynamics underlying a long-term associative memory,” *Nature*, vol. 543, no. 7647, Art. no. 7647, Mar. 2017, doi: 10.1038/nature21682.
[79] J. C. Magee and C. Grienberger, “Synaptic Plasticity Forms and Functions,” *Annu. Rev. Neurosci.*, vol. 43, no. 1, pp. 95–117, 2020, doi: 10.1146/annurev-neuro-090919-022842.
[80] P. Mishra and R. Narayanan, “Stable continual learning through structured multiscale plasticity manifolds,” *Curr. Opin. Neurobiol.*, vol. 70, pp. 51–63, Aug. 2021, doi: 10.1016/j.conb.2021.07.009.
[81] C. R. Lee, L. Najafizadeh, and D. J. Margolis, “Investigating learning-related neural circuitry with chronic in vivo optical imaging,” *Brain Struct. Funct.*, vol. 225, no. 2, pp. 467–480, Mar. 2020, doi:

10.1007/s00429-019-02001-9.
[82] N. A. Steinmetz *et al.*, “Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings,” *Science*, vol. 372, no. 6539, p. eabf4588, Apr. 2021, doi: 10.1126/science.abf4588.
[83] S. Zhao *et al.*, “Tracking neural activity from the same cells during the entire adult life of mice,” *Nat. Neurosci.*, pp. 1–15, Feb. 2023, doi: 10.1038/s41593-023-01267-x.
[84] M. Dabagia, K. P. Kording, and E. L. Dyer, “Comparing high-dimensional neural recordings by aligning their low-dimensional latent representations.” arXiv, May 17, 2022. doi: 10.48550/arXiv.2205.08413.
[85] A. H. Williams, E. Kunz, S. Kornblith, and S. Linderman, “Generalized Shape Metrics on Neural Representations,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, pp. 4738–4750. Accessed: Apr. 27, 2022. [Online]. Available: https://papers.nips.cc/paper/2021/hash/252a3dbaeb32e7690242ad3b556e626b-Abstract.html
[86] A. Pellegrino, H. Stein, and N. A. Cayco-Gajic, “Disentangling Mixed Classes of Covariability in Large-Scale Neural Data.” bioRxiv, p. 2023.03.01.530616, Mar. 02, 2023. doi: 10.1101/2023.03.01.530616.
[87] A. H. Williams *et al.*, “Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis,” *Neuron*, vol. 98, no. 6, pp. 1099-1115.e8, Jun. 2018, doi: 10.1016/j.neuron.2018.05.015.
[88] A. R. Galgali, M. Sahani, and V. Mante, “Residual dynamics resolves recurrent contributions to neural computation,” *Nat. Neurosci.*, vol. 26, no. 2, Art. no. 2, Feb. 2023, doi: 10.1038/s41593-022-01230-2.
[89] M. G. Perich *et al.*, “Inferring brain-wide interactions using data-constrained recurrent neural network models.” bioRxiv, p. 2020.12.18.423348, Mar. 11, 2021. doi: 10.1101/2020.12.18.423348.
[90] N. A. Roy, J. H. Bak, A. Akrami, C. D. Brody, and J. W. Pillow, “Extracting the dynamics of behavior in sensory decision-making experiments,” *Neuron*, vol. 109, no. 4, pp. 597-610.e6, Feb. 2021, doi: 10.1016/j.neuron.2020.12.004.
[91] K. V. Kuchibhotla *et al.*, “Dissociating task acquisition from expression during learning reveals latent knowledge,” *Nat. Commun.*, vol. 10, no. 1, p. 2151, May 2019, doi: 10.1038/s41467-019-10089-0.
[92] B. A. Richards *et al.*, “A deep learning framework for neuroscience,” *Nat. Neurosci.*, vol. 22, no. 11, Art. no. 11, Nov. 2019, doi: 10.1038/s41593-019-0520-2.
[93] B. Bordelon and C. Pehlevan, “The Influence of Learning Rule on Representation Dynamics in Wide Neural Networks.” arXiv, Oct. 05, 2022. doi: 10.48550/arXiv.2210.02157.
[94] Y. Cao, C. Summerfield, and A. Saxe, “Characterizing emergent representations in a space of candidate learning rules for deep networks,” in *Advances in Neural Information Processing Systems*, 2020, pp. 8660–8670. Accessed: Apr. 12, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6275d7071d005260ab9d0766d6df1145-Abstract.html

Papers of special interest (*)

- [47] Failor et al, 2021
  Using longitudinal imaging in mouse V1 during a visuomotor association task, the authors show that stimulus discriminability may not necessarily be the driver of learning-related changes in sensory areas, instead proposing that orthogonalisation of population responses to task-relevant stimuli could enable dissociable sensorimotor associations. They further link this change in geometry to cellular mechanisms that tend to sparsen population codes.

- [56] Finkelstein et al, 2021
  The authors study neural and behavioral responses in mice that are trained to perform a decision-making task, where they may experience distractor stimuli during the delay period. By analyzing task-trained RNNs performing the same task, they argue that exposure to the distractor reshapes recurrent dynamics so as to deepen the attractor wells corresponding to different choices, increasing the robustness to perturbations for the distractor-trained mice.

- [44] Libby and Buschman, 2021
  The authors examine how A1 population activity evolved during implicit statistical learning of auditory sequences. By studying the geometry of stimulus representations, they proposed that while learning could lead to an alignment of sensory encoding axes for associated stimuli, a separate orthogonal "memory" subspace maintained the identity of recent stimuli, potentially allowing downstream readout of current and recent sensory stimuli without interference.

- [73] Losey et al, 2022
  This study examines how neural activity is reorganized to enable multiple behaviors to be learnt without interference. In a BCI learning paradigm, the authors found that when animals return to a familiar map after learning a novel BCI map, neural activity had primarily shifted along dimensions that did not affect behavioral performance with the familiar map, while also reflecting a memory trace corresponding to the novel map.

- [62] Driscoll et al, 2022
  The authors demonstrated that training a single RNN on 15-20 common tasks produced a compositional code: tasks could be structured as modular computations implemented via dynamical building blocks such as attractors, decision boundaries and bifurcations, that were shared across sub-tasks with similar computations. They also suggest that these modular dynamical motifs can be recombined for fast learning of new tasks (transfer learning).

Papers of outstanding interest (**)

- [50] Chadwick et al, 2023
  The authors propose a new theory that the slow and fast modes of recurrent circuit dynamics realign over perceptual learning to aid integration of noisy sensory evidence. They show that V1 population activity in mice learning to perform a sensory discrimination task is more consistent with their dynamical realignment hypothesis than by a simple slowing of timescales in the dynamics.

- [35] Farrell et al., 2022
  The authors show that RNNs learning under gradient descent tend to lead to low-dimensional representations that support generalizations, sometimes following a period of temporary expansion. They further pinpoint noise in weight updates as a key driver of dimensionality compression, suggesting a possible role of synaptic noise on the dimensionality of learned neural representations.

- [74] Latimer and Freedman, 2023
  This study provides an indication of how training history (or the specific sequence of tasks learnt) can affect the computational strategies that animals use for solving the tasks, as reflected in both the low-dimensional geometry of neural representations and the pattern of behavioral errors.

- [54] Sun et al., 2022
  The authors ask how preparatory activity in the primate motor cortex changes throughout learning of a sequence of motor adaptation tasks. They find structured changes in the geometry of preparatory states, including task-specific shifts along a specific dimension consistent with an index of different motor tasks. Importantly, even after washout, the learned representations retained information about previously learned tasks.

- [63] Goudar et al., 2023
  By training RNNs to learn a series of sensorimotor mappings, the authors studied the formation of low-dimensional representations which were reused across mappings, and which increasingly encoded an abstraction of the shared task structure. Using novel analyses to dissect the relative contributions of using different local dynamics versus reshaping dynamics through recurrent plasticity, they showed how this reuse led to increased learning efficiency and mitigated ‘forgetting’.

Figures

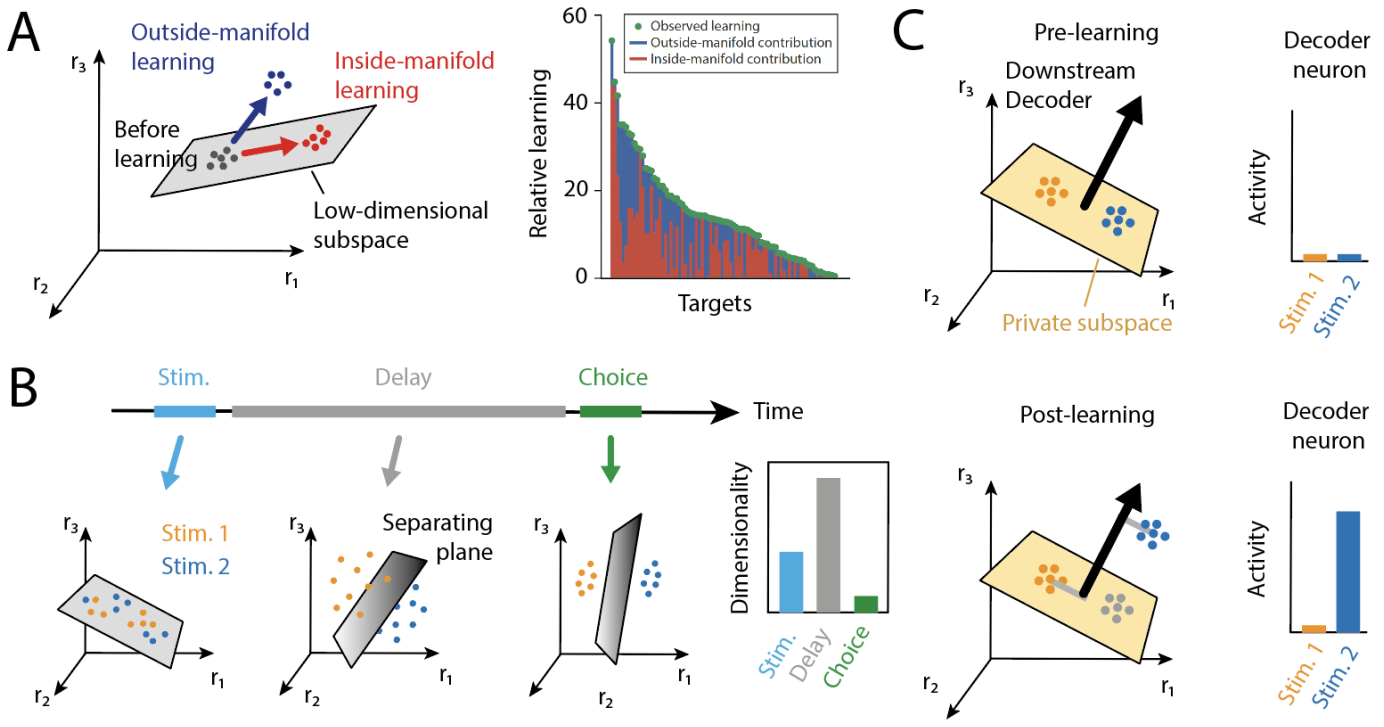

Figure 1: Learning-induced changes in manifold geometry.

(A) Left: Neural activity is often observed to lie in a low dimensional subspace (i.e., linear manifold) (gray). The effect of learning on manifold structure can be decomposed into inside-manifold learning (red) or outside-manifold learning (blue) contributions, depending on changes in activity. Right: Level of learning in a BCI task (green dot, last day progress relative to first day) sorted for different cursor targets. For each target, relative learning level was decomposed into the change in performance attributed to inside-manifold or outside-manifold contributions (see [20] for details). Both effects were found to contribute to learning, suggesting neural manifolds can change over long-timescale learning. Adapted from [20]. **(B)** The dimensionality of learned neural representations depends on the task period. In an ANN trained on a classification task, the dimensionality of each stimulus representation increases transiently to improve linear discriminability, but becomes more compact during the choice period to enable a robust downstream readout. In this scenario, learning could simultaneously increase delay-period dimensionality while decreasing choice-period dimensionality. Adapted from [35]. **(C)** Example of re-orientation of task relevant dimensions to a decoder axis. Top: (Pre-learning) Representations of two stimuli are separable but in the same decoder-null subspace that is orthogonal to the dimension representing the decoding axis (left). Both stimuli produce weak responses in the decoder neuron (right). Bottom: After training, the representation of Stim. 2 shifts along the decoding axis, producing a larger projection onto the decoding axis (left) and hence a larger downstream response (right). More generally, network activity can shift between a decoder-null or decoder-potent subspace to suppress or enhance the readout. See also [46].

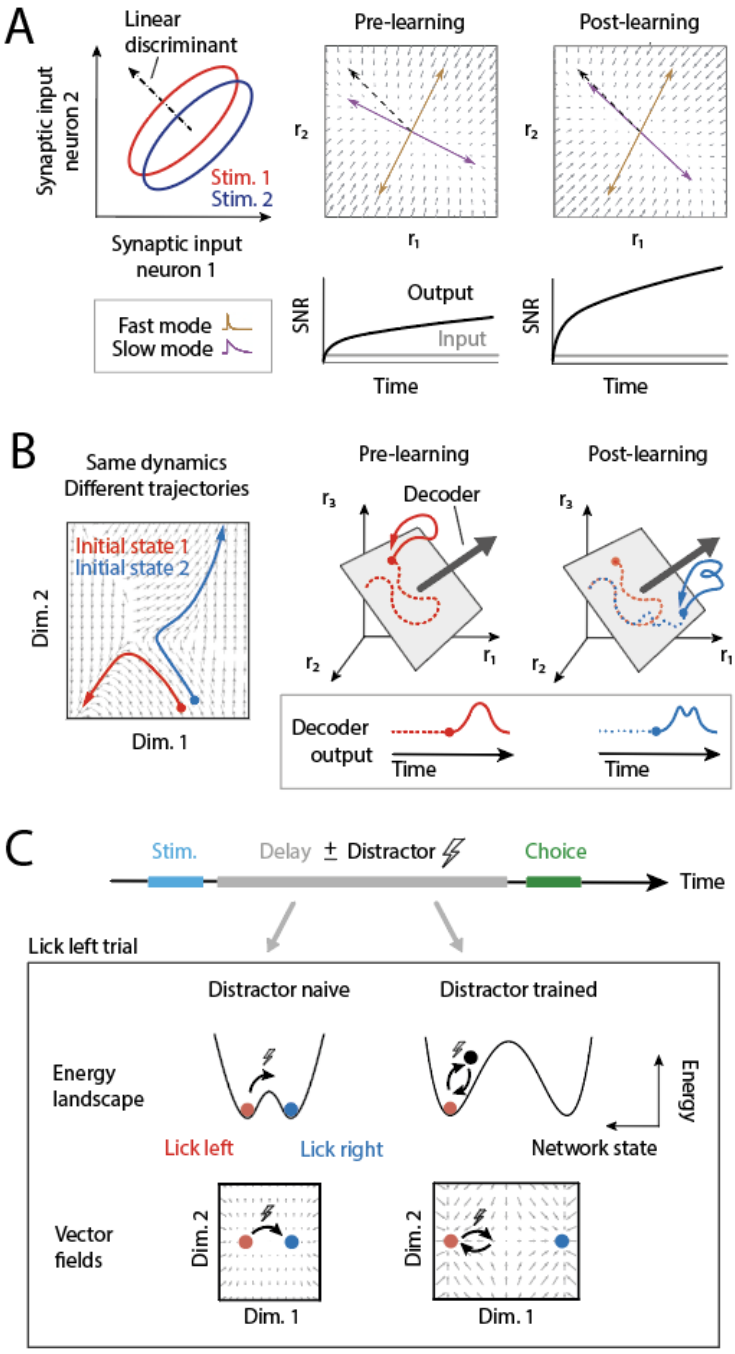

Figure 2: Learning-induced changes in dynamics.

(A) Alignment of slow network dynamics with informative input modes. The network learns to discriminate between two stimuli by performing leaky integration of noisy feedforward input. Left: Distributions of synaptic input onto two neurons for two stimuli at any given time. The linear discriminant axis separates the two representations (dashed black) and is the most informative input direction. Middle: Recurrent dynamics illustrated by a vector field diagram. The gray arrows indicate the (instantaneous) velocity of changes of the system state (here, neural population activity). In this example, recurrent dynamics are decomposed into a slowly-decaying (purple) and a fast-decaying (orange) mode (top). Stimulus information (or SNR; signal-to-noise) in network output grows as compared to input. Right: Over training, realignment of the slow mode to the input linear discriminant axis means stimulus information decays less quickly during integration, leading to more rapid evidence accumulation. Adapted from [50]. **(B)** Changes in initial conditions can trigger

different neural trajectories. Left: Example vector field visualized along two dimensions of neural activity space. By changing the initial network activity state (red and blue dots), the same underlying dynamical structure can be used to produce different neural trajectories, which may have different consequences for downstream decoding. Right: Motor planning-related activity (dashed red) is orthogonal to the movement-decoding axis and can therefore set the initial condition (red circle) without producing any movement. Upon receiving a 'Go-cue', the network activity evolves as the trajectory in red and produces motor output. After training, the network follows a different preparatory trajectory (dashed blue) leading to a new initial condition (blue circle) at the end of motor planning. This triggers a new trajectory (blue) that changes the motor output. **(C)** Changes in attractor dynamics during a stimulus detection task. Top: Mice were trained to report the presence (lick right) or absence (lick left) of a stimulus; during the cue, a distractor stimulus (lightning bolt) was sometimes added as a perturbation. Bottom: Schematic of the attractor landscape and corresponding vector field in distractor-naive (left) and distractor-trained (right) mice. During the delay period, network activity is in one of the attractor basins, maintaining a motor plan. In distractor-naive mice, distractors are able to switch the network activity from one attractor to another. In distractor-trained mice, the network state is more robust due to greater separation of attractor basins. Adapted from [56].

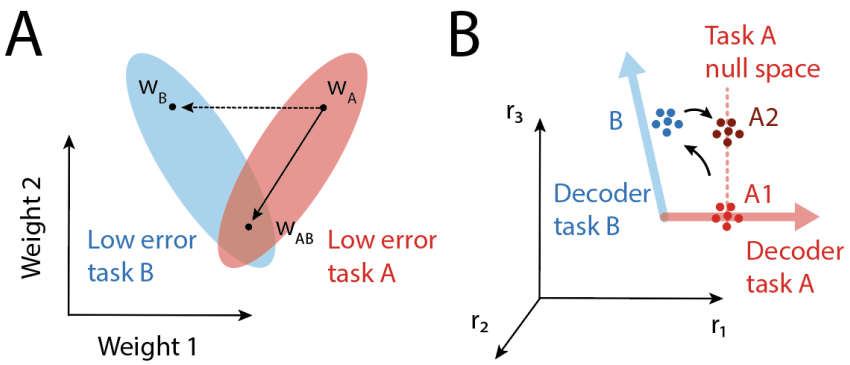

Figure 3: Sequential skill acquisition depends on previously acquired skills.
(A) Continual learning algorithms provide a solution to catastrophic forgetting by constraining weight changes so that they do not lead to poor performance in previously-learnt tasks. In this example, an ANN that has been pre-trained on task A (initial weights w_A) lies in the red oval corresponding to solutions with low task A error. When trained on task B, standard gradient descent leads to weights w_B, which have low error on task B (blue oval) but poor performance on task A ("forgetting"; dotted line). Instead, continual learning updates the weights so that the resulting network can perform both tasks (w_{AB}). Figure adapted from [71] and [67]. **(B)** Schematic of learning in subjects trained sequentially on two BCI tasks A and B (determined by red and blue decoder axes). Over learning, neural representations are updated from having high projection on decoder A to having high projection on decoder B. Interestingly, when subjects are then re-exposed to task A (A2; maroon), the representations lie in a region corresponding to high performance in both tasks, revealing a 'memory trace' of map B. Adapted from [73].