Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Neurobiology

From lazy to rich to exclusive task representations in neural networks and neural codes

Matthew Farrell[1], Stefano Recanatesi[2] and Eric Shea-Brown[2]

Abstract

Neural circuits—both in the brain and in “artificial” neural network models—learn to solve a remarkable variety of tasks, and there is a great current opportunity to use neural networks as models for brain function. Key to this endeavor is the ability to characterize the *representations* formed by both artificial and biological brains. Here, we investigate this potential through the lens of recently developing theory that characterizes neural networks as “lazy” or “rich” depending on the approach they use to solve tasks: lazy networks solve tasks by making small changes in connectivity, while rich networks solve tasks by significantly modifying weights throughout the network (including “hidden layers”). We further elucidate rich networks through the lens of compression and “neural collapse”, ideas that have recently been of significant interest to neuroscience and machine learning. We then show how these ideas apply to a domain of increasing importance to both fields: extracting latent structures through self-supervised learning.

Addresses

[1] John A. Paulson School of Engineering and Applied Sciences, Harvard University and Center for Brain Science, Harvard University, United States

[2] Applied Mathematics, Physiology and Biophysics, and Computational Neuroscience Center, University of Washington, United States

Corresponding author: Shea-Brown, Eric (etsb@uw.edu)

Current Opinion in Neurobiology 2023, **83**:102780

This review comes from a themed issue on **Computational Neuroscience 2023**

Edited by **Jeanette Hellgreny Kotaleski** and **Tatjana Tchumatchenko**

For a complete overview see the Issue and the Editorial

Available online 25 September 2023

https://doi.org/10.1016/j.conb.2023.102780

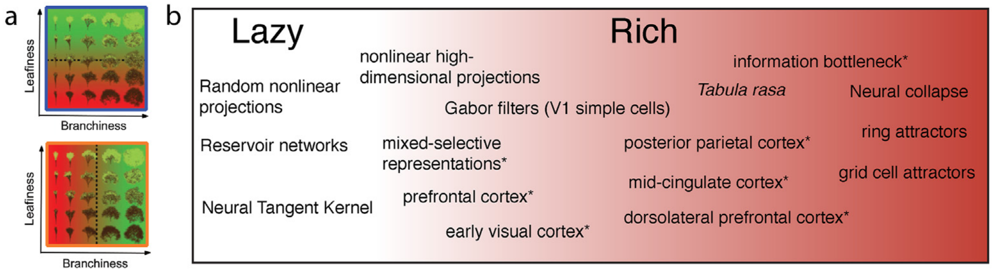
Introduction

When we learn and develop, from learning to play chess to learning to walk to learning a relatively controlled laboratory task, the brain undergoes changes that specialize neural circuits to certain functions. However, the degree of specialization, as well as the elements of the environment and task that are specialized for, varies significantly across the brain. There is evidence that some brain areas hold a potentially high-dimensional (high-d), possibly random mix of many sensory features and task variables, as in high-d mixed-selective representations [1,2]. There is also evidence for other brain areas holding information that is very exclusively focused on task variables, such as category identity in classification tasks [3]. However, a unifying perspective that explains these phenomena is still developing, and it is not clear when and where general-purpose, task-agnostic representations should be found versus exclusive, task-specific representations, and what the implications may be.

Artificial neural networks are proving a remarkably useful model system for addressing such questions [4]. In this approach, a neural network is trained to perform a task comparable to that performed by an animal in an experiment; the network is then analyzed to gain insight into plausible neural mechanisms and representations that support task execution. An obvious but appealing aspect of this approach is that representations in the artificial system can be measured with perfect completeness and precision. By studying the representations formed by neural networks, we can probe the functional role of representational structure. Moreover, by efficiently simulating and comparing representations that arise in many different settings, or performing mathematical analysis of well-defined learning processes, we can ask explicitly whether and how structure in these representations depends on the underlying tasks, as well as on varied assumptions about the networks and their learning rules. Of course, how much this informs the underlying biology is a deep, classic [5] and still open question, but a still-growing body of recent evidence underscores that the underlying representational principles may as well be at work in animal brains [6–8].

A striking finding that has recently emerged from theoretical investigations of neural networks is the large diversity of qualitatively different solutions these networks find, depending on initialization scheme, optimization procedure, and other details. These results belong to a sub-field that is often referred to as “feature learning.” A feature is simply an aspect of inputs in a particular domain; feature learning describes how learning systems might access or extract useful features that support performing tasks within this domain. In neuroscience, the closely related concept of a neural representation is more familiar, which ultimately refers

Figure 1



Current Opinion in Neurobiology

(**a**) Depiction of a context-dependent classification task, taken from Ref. [9]. Top: task is to classify images based on leafiness. Bottom: task is to classify based on branchiness. (**b**) An idealized schematic of the lazy to rich spectrum of representations that can follow from task learning. Asterisks denote items that are speculative or do not fit neatly into one place.

to how the biological network represents specific aspects of the external world. Often it is the case that useful features are encoded in neural representations. For instance, edges are useful features of images that support image classification; it is plausibly for this reason that edge-detecting simple cells are often found in mammalian visual cortex.

In this article, we highlight an intriguing set of recent findings about learned representations in artificial neural networks, and how they may shed light on biological neural representations and the underlying learning processes. In particular, we study:

1. When are networks *lazy*, in that they learn to accomplish tasks without changing their representations? When are networks instead *rich*, in that they change their representations over the course of learning?
2. In the case of rich networks, when does this richness reach the level of being *exclusive*, where their representations not just learn but also isolate task-relevant information?
3. How do exclusive networks shape representations in self-supervised tasks, such as predictive learning, to uncover hidden task structures or variables?

The third question has rapidly become of high importance to machine learning, as training regimens for artificial neural networks are increasingly dominated by a self-supervised initial stage, as well as in neuroscience, as self-supervised tasks are increasingly used as models to explain marquee neural representations such as place and grid cells.

Reviewing these topics will carry us across a spectrum from less to more extreme examples of feature learning: from no feature learning at all to exclusive feature learning that actively removes from representations any information inessential to the task at hand. Figure 1a illustrates the concept of task-relevant and task-irrelevant features by showing a task where images of trees are classified either according to the tree leafiness or branchiness, according to a context signal. In the first context, aspects of the branchiness features are irrelevant to the leafiness features, as well as vice-versa in the second. If both contexts are used through training, both aspects are important. Figure 1b gives a rough schematic of the range of lazy to rich to exclusive learning by placing models, concepts, and brain areas along this spectrum.

Below, we begin with the regime in which features are not learned, known as the “lazy” regime in the feature learning literature. This regime has been important in the development of mathematical theories predicting the behavior of neural networks [10–13]. We then move to the regime in which feature learning occurs, known as the “rich” regime. Within the rich regime, we will investigate the different degrees of richness that can occur, which can be measured by the extent to which networks learn to exclusively represent information required for a specific task while rejecting other incoming information.[1] Throughout, we highlight insights that underlying concepts can provide for neurobiology (see e.g., Refs. [9,14,15]).

The lazy learning limit: learning tasks without encoding them

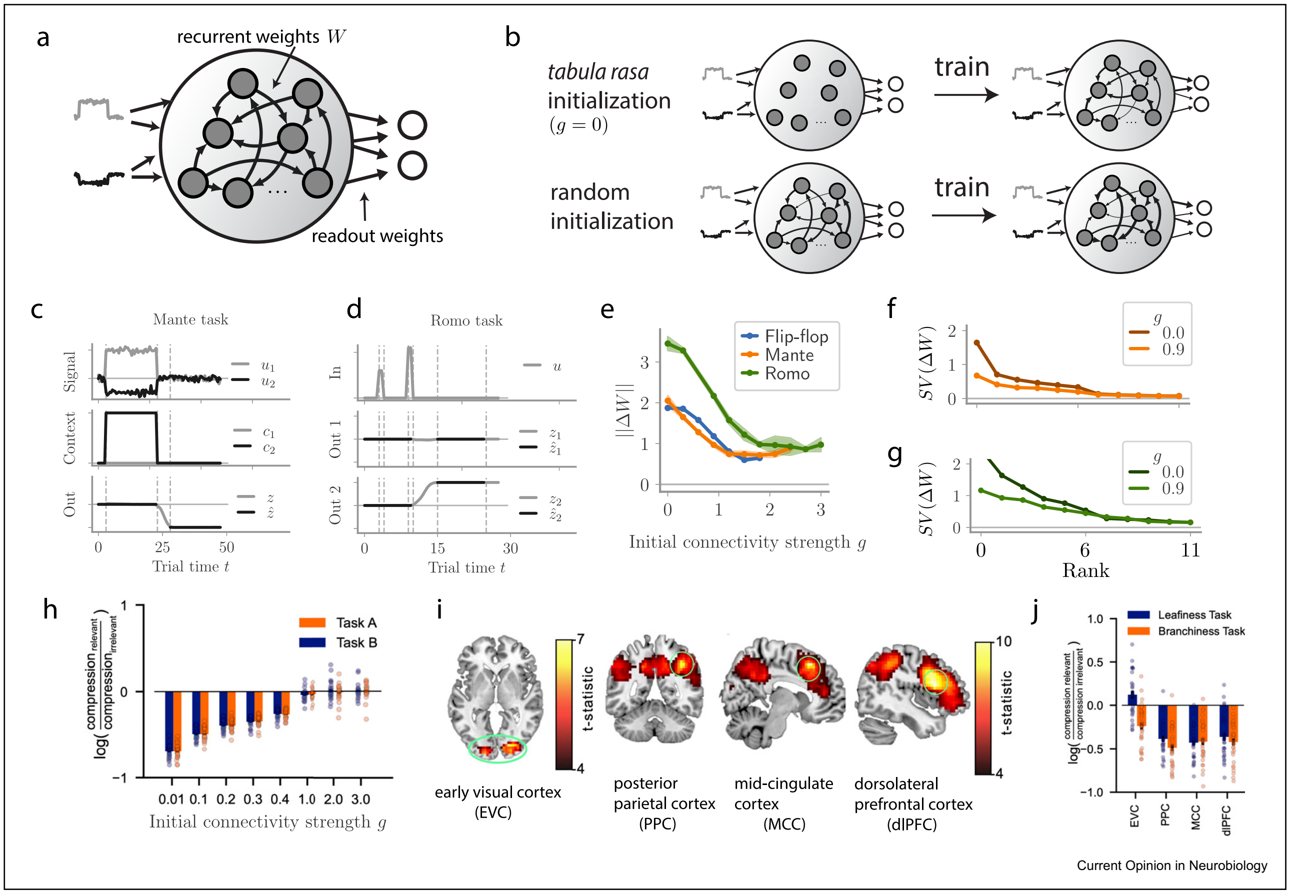
Lazy learning describes a lack of task-relevant changes in the representation of a neural network (see below for a more formal definition). The most basic way in which lazy learning occurs is in model networks where internal connection weights are simply held constant by design: they are never allowed to vary from their initial values, which are chosen before learning begins. In this case, the only connection weights that are changed during

---

[1] Note that this terminology can be confusing, as the “rich” regime can involve *removing* information. Here richness refers to whether or not network representations are “enriched” (i.e. modified) over training in a way that reflects desired task outputs.

learning are the output weights linking a network's internal activity to its final response. As a prominent example, networks whose internal connections are initialized with random internal connection weights may perform a sufficiently large number of transformations of their inputs in their internal layers (e.g., random nonlinear projections to a high-d space), so that the networks can solve many tasks by changes in readout weights alone. For example, in the illustration of Figure 2a, if the network is large, its neurons are nonlinear, and the internal (recurrent) weights W are sufficiently strong and random, then a vast set of input—output maps may be created by leaving W fixed and just tuning the readout weights. This is the approach taken by the related frameworks of support vector machines [16], random feature models [17], kernel machines [18], reservoir computers [19,20], Koopman operators [21], and neural network Gaussian processes [22–25]. In particular, reservoir computers have featured prominently in the development of theories of neural

Figure 2



When do neural networks learn to represent task structure? Illustrating the influence of initial connectivity. (**a**) Illustration of a recurrent neural network. Inputs arrive and are transformed in internal, or "hidden," layer(s): here, this is the recurrent layer in the central circle. This internal representation is then read out, via readout weights, to become the output for the task. (**b**) Cartoon comparing training from a *tabula rasa* initialization and a random initialization. Initial connectivity strength is denoted by g. *Tabula rasa* in this case corresponds to $g = 0$. Panels (**c**)–(**g**) taken from Ref. [30]. (**c**) Depiction of an instance of the "Mante task." Here, the network receives two noisy input signals, one of which is relevant in a given trial of the task and the other of which serves as a distractor. The network also receives "context" inputs indicating the relevant signal. The network is tasked with outputting whether the relevant signal was positive or negative on average over the time course of the trial. (**d**) Depiction of an instance of the "Romo task." Here, the network receives two input pulses separated by a delay, after which the network is tasked with outputting which of the two pulses had the larger amplitude. (**e**) Frobenius norm of the change in weights resulting from training, as a function of initial connectivity strength. Colors denote different tasks. (**f**) First 11 singular values of the change in weights resulting from training on the Mante task. Color denotes initial connectivity strength. (**g**) As in (**f**), but for the Romo task. (**h**)–(**j**) Adapted from Ref. [9]. See Ref. [9] for details. (**h**) Compression of task-irrelevant versus task-relevant information in a neural network with a single hidden layer trained on a simplified version of the context-modulated classification task illustrated in Figure 1a, as a function of the initial connectivity strength g. More negative values indicate stronger compression of task-irrelevant information which is indicative of rich feature learning. (**i**) Representation analysis of BOLD signals during human execution of the context-modulated classification task illustrated in Figure 1a. Left panel: similarity of BOLD signals with input-specific features (signifying a lazy representation). Right three panels: similarity of BOLD signals with output/choice-specific features (signifying a rich representation). (**j**) As in (**h**), but measured for activity (representations) in different human brain regions during the execution of the context-modulated classification task illustrated in Figure 1a.

computation [26]. This perspective of random nonlinear projections followed by learned readout weights is also prominent in circuit models for early sensory and cerebellar processing, for example, Refs. [9,27–29].

Perhaps more surprisingly, even when we do allow synapses to be modified in both internal and readout layers (i.e., training the network "end-to-end"), lazy learning can still occur (cf. [11]). This phenomenon has come under intense focus in recent years. While networks will typically engage in some amount of task-relevant feature learning, it is theoretically convenient to consider limits in which feature learning either does or does not occur. One such limit is called the neural tangent kernel (NTK) limit. NTK theory says that if neural networks are initialized in a certain way, and if the size of each layer in the network is taken to infinity (as for neural circuits with vast numbers of internal neurons), then the evolution of the network outputs through training can be described by a relatively simple mathematical object; namely, a linear system of ordinary differential equations. The coefficient matrix of this system is the NTK. This theory is enticing because the simple training dynamics allow theoreticians to analytically predict aspects of the behavior of the network through training, such as the error the network will have on held-out data after training a certain number of steps [54].

These concepts have been applied to many network models, including convolutional neural networks as well as recurrent neural networks [31,32]; however, simple unstructured feedforward networks are currently the best understood. Early evidence for the existence of a lazy regime for recurrent networks was also found in Ref. [33].

In the NTK limit, the network before training is randomly initialized, and the network weights after training remain close to this random initialization [34], such that the intermediate representations are still essentially random [11]. For this reason, networks initialized according to the NTK theory are said to be in the lazy regime. Due to this lack of feature learning, networks in the lazy regime resemble the support vector machines, random feature models, and other related ideas described above. A precise definition of the lazy regime can be found in Ref. [11]; for our purposes, it is sufficient to understand that the intermediate feature representations do not change in a task-relevant way.[2] This said, many adjustments to the initialization scheme and other details can result in more substantial learning in internal representations [11]; in general, much remains to be discovered (see Refs. [11–13,35–41]).

Interpretations for neurobiology

We pause to highlight three points important for relating the lazy regime to neurobiology. First, we reiterate that the scaling limits considered in these works are motivated mathematically, as they allow for a clean theory to be built, but that the ideas of rich and lazy can be used in a less formal way and applied to neural networks that are not infinite in size. In addition, while the formal theory as we have introduced it above refers to entire networks as potentially being lazy, we can use this concept for individual layers or individual representations. Finally, we note that while lazy learning is defined above in terms of network initialization, there are multiple ways that the concept of initialization could apply in biological settings; we discuss some possibilities in the section "Interpretations for neurobiology, revisited" below.

Leaving room for learning: Rich learning of neural representations

Above, we reviewed how networks that start with strong internal network weights can show lazy learning, in which no meaningful learning of task structure occurs in a network's internal representations. The theoretical work underlying this phenomenon also suggests how networks may move away from this lazy learning as the (relative) strength of initial weights is tuned down (e.g., see Ref. [11]). Here, we highlight some recent work from computational neuroscience that quantifies the type of learning that occurs in this setting.

Rich learning from scratch: the *tabula rasa* regime

The limit of taking very small initial weights is called the *tabula rasa* regime and has played an important role in our understanding of network function. Here, the opposite of lazy behavior—termed *rich learning*—occurs: the task structure is strongly learned and represented in the underlying neural network. Figure 2b shows an illustration that contrasts *tabula rasa* versus the strong and random initialization schemes similar to the NTK initialization discussed above. The behavior of the *tabula rasa* regime was made explicit in elegant mathematical studies of deep linear networks [36,42–44]. In this setting, the network picks up the structure of the task in a parsimonious fashion through training, with the modes of the input–output covariance matrix (the principal components) being transferred to the weights in order of their magnitude. In this way, network weights and activities clearly represent the structure of the task. With *tabula rasa* initialization, at least in the tractable case of linear feedforward networks, this is the only structure that weights represent.

Titrating away from *tabula rasa*

A recent study focusing on recurrent neural networks explores the intermediate ground between larger and smaller weight initializations [30]. In this work, the

---

[2] To be a bit more precise, the training time would need to diverge to infinity along with the size of the network in order for features to be learned.

tabula rasa behavior of the network is reconciled with a non-vanishing random initialization. This is done in the context of three tasks, two of which are shown in Figure 2c and 2d. A key result is how random components of network weights present at initialization perturb the learning dynamics away from *tabula rasa* behavior. In particular, this random initialization is "sticky", with the network weights retaining higher-rank components through training (Figure 2e–2g). In this case, the network after training may assume a lower-dimensional structure more dominated by a single mode if the random component is small or a higher-dimensional structure if the random component is large at initialization (Figure 2f and 2g). In general, the main story that emerges seems to be that *changes* induced by learning have a rank that matches that of the task; often, this is low-rank. See also Refs. [12,36,41,45,46] for explorations beyond the *tabula rasa* regime, which find similar principles at work.

In the cases explored in these studies, random components in the weights present at initialization tend to remain throughout training. In the section "Learning to be rich and exclusive" below, we will study the even more dramatic case of active compression of task-irrelevant information, where the random components present at initialization are significantly reduced through training.

Interpretations for neurobiology, revisited

We reiterate that both lazy and rich learning are defined above in terms of network initialization and how much the representation changes from this initialization. In biological circuits, what an initial network means is somewhat up to interpretation. For instance, a neural circuit may be considered *tabula rasa* at the outset of development (connections between neurons being weak or non-existent), and the learning process could be a mix of genetically determined development along with synaptic modifications driven by experience in the world. Conversely, circuits that engage in learning a new task after a lifetime of learning other tasks may be considered to have stronger initial weights. Most settings will, of course, lie in between these extremes, and a great deal about the factors that control rich vs. lazy learning outcomes doubtless remains to be discovered. In the meantime, the spectrum of representations that can arise in neural networks with the types of initializations studied to date forms intriguing and testable hypotheses for experiments, as we review next.

Connections to experiments

Theoretical work exploring lazy and rich regimes has counterparts in experimental neuroscience studies such as [9]. In this study, the authors design a context-dependent task (illustrated in Figure 1a) that can be solved in two ways: random projections to a high-d space followed by a trained readout (corresponding to the lazy regime), or using intermediate weights to transform the representation to a compact form (corresponding to the rich regime). In line with the computational work reviewed above, the authors found that a neural network trained on a slightly simplified version of the task would take on the rich or lazy solution depending on the strength of connection weights in its (random) initialization (see Figure 2h). They then studied neural activity in the brains of humans and macaque monkeys during performance of this task and found that early sensory brain areas represent information in a way consistent with the lazy regime, whereas higher-order areas such as posterior parietal cortex have representations more consistent with the rich regime (Figure 2i and 2j).

Much work remains in charting out the strategies used in different brain areas across a range of tasks, and across different levels of functional hierarchy. To help address the need to understand brain representations across a large range of tasks, the authors of [47] compare across-task similarity matrices of brain representations to those formed by neural networks in the rich and lazy regimes and find closer matches with rich networks. The study [48] also observes that deeper brain areas have a rich-learning-like representation and investigates the generalization properties of these representations. These studies indicate the power and importance of the theory of rich and lazy networks when modeling neural representations.

Allied and very interesting concepts appear in the form of "mixed-selective" representations as observed in monkey prefrontal cortex [1,2], which correspond to a high-d nonlinear mapping of sensory and task variables into the neural representation. We note that a mixed-selective representation is not necessarily equivalent to a lazy representation—the degree to which the representation is lazy depends on the degree to which the nonlinear mapping of input and task data is specifically tailored to the relevant tasks—a fully lazy representation will instead be random (see Figure 1b for concrete examples). It is an interesting future direction to ascertain the degree to which mixed-selective and other neural representations are indeed random (see for example Refs. [49–51]) and to explore connections to mechanisms underlying lazy learning.

Learning to be rich and exclusive: compressing away task-irrelevant information

When networks operate outside the lazy regime, they change their internal representations in accordance with task demands. But do they become single-minded in this regard, compressing away input and information that is not directly relevant to the task at hand? This is a very strong way in which networks could encode tasks, since it isolates only the task-relevant information. It can involve the active removal of any information initially present but irrelevant to the task; signatures of this removal should be measurable in both experiments and model simulations. Such exclusive representations can also, in principle, have significant functional implications, such as enabling fast downstream learning and generalization on similar tasks [2,52–54], while limiting the ability of downstream networks to learn new tasks that require information that has been discarded [1,55].

Two complementary perspectives

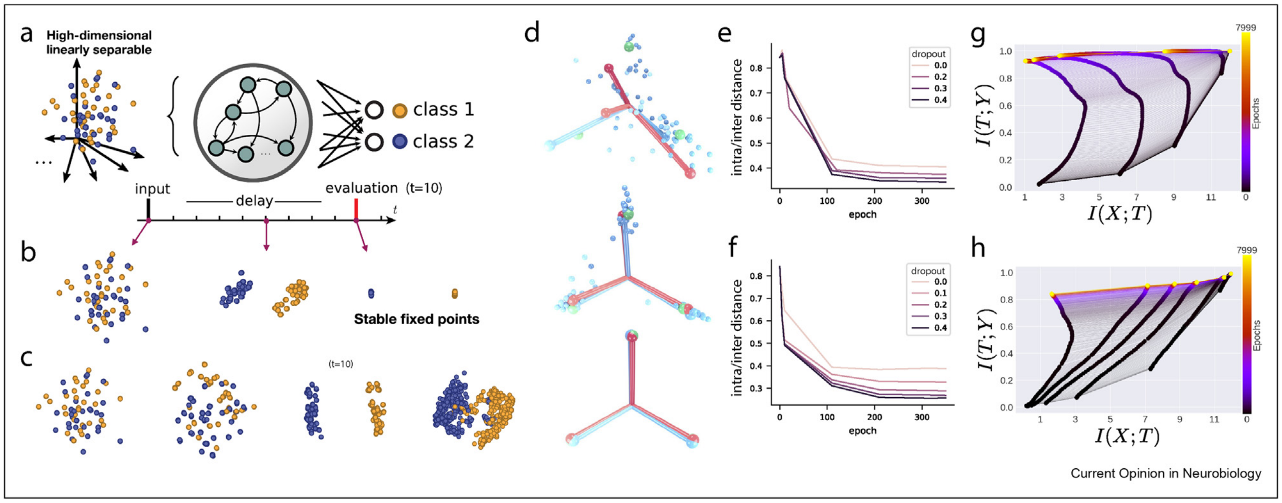
We next review two perspectives on the compression of task-irrelevant information from the recent neural network literature. We first describe neural collapse, as it takes a direct "geometric" description of this compression that is easy to visualize. We then return to the earlier, as well as inspirational, idea of the information bottleneck, which quantifies the compression of task-irrelevant inputs using mutual information.

Dimension compression and neural collapse

Recent studies [41,53,56–60] have taken a geometrical view on how networks can learn to actively compress away aspects of inputs that are not directly relevant to the task at hand, a phenomenon elegantly described by Papyan, Donoho, and colleagues as "*neural collapse*" [57]. Here, the structure of the recurrent neural network activity at later timesteps—or of deep neural network activity in penultimate layers—becomes very low-dimensional, in certain cases even collapsing to a set of single points. This occurs even though the network is initialized with weights that form high-d representations carrying both task-relevant and task-irrelevant information.

In particular, for tasks with discrete outputs (or categories) into which inputs are grouped, the representation after training can become highly compressed around each category [53,57,60]. This behavior was observed in the context of change detection tasks [60] and discrete classification tasks [53], where it is referred to as dimensionality compression. Figure 3a and b, modified from Ref. [53], gives an illustration of this phenomenon. Here, the task is delayed classification. First, an input arrives at timestep zero. A recurrent neural network then processes this input for some number of timesteps, until an evaluation time when the network activity is read out. Figure 3b shows the phenomenon at hand: a network that has learned this task strongly compresses its inputs into clusters corresponding to the task categories. Here, the

Figure 3



Current Opinion in Neurobiology

When do neural networks compress task-irrelevant inputs and information? Illustrating neural collapse and the information bottleneck. (**a**) Illustration of a recurrent network solving a delayed classification task with high-d, and hence linearly separable, inputs; see text. The network's cost function is defined via categorical cross-entropy, with no additional penalty terms added. (**b**) Visualization of the trained network representation's top two principal components as it evolves through time. (**c**) As in (**b**), but for stronger initial connectivity strength. (**d**) Illustration of the representation of the penultimate layer of the convolutional neural network VGG13 trained on CIFAR10 (adapted from Ref. [57]), projected onto the top three principle component axes (axes not shown). Green spheres denote axes of a two-dimensional simplex; red balls and sticks represent the normal vectors for classifying hyperplanes; blue balls and sticks denote class means; and small blue spheres represent penultimate layer features. Top is before training, middle is at an intermediate stage of training, and bottom is after training. Note that this representation is induced by the training set, not the test set. (**e**)–(**f**) Plots of average intra-class divided by inter-class distances over training (epochs) in the penultimate layer of Resnet18, where dropout is applied during training (but not during the evaluation of distances). Lower values indicate higher compression of task-irrelevant information. Loss used is mean squared error. Shading denotes amount of dropout applied (probability that a neuron's output will be silenced). (**e**) Optimizer used is stochastic gradient descent with momentum value of 0.9, a common optimizer choice in training neural networks. (**f**) Optimizer is RMSprop without momentum, another common choice. (**g**) Mutual information between a convolutional neural network layer's representation T and the inputs X (x-axis) as well as the outputs Y (y-axis) in a convolutional neural network with tanh nonlinearities (taken from Refs. [52,61]). Each curve corresponds with a different layer T, and color denotes training time (epochs). (**h**) As in (**g**), but with ReLU nonlinearities (taken from Ref. [61]).

network at the initialization of learning approximately preserves the structure of the inputs; hence, arriving at a collapsed representation requires all of the task-irrelevant structure of the inputs to be quashed through learning [53]. Note that in this example, the two categories of inputs are linearly separable in the input space. Thus, the inputs do not need to be reformatted by the recurrent network in order to solve the task, as they could be classified with 100% accuracy by the output weights alone. Rather, the highly structured representations that form are an interesting by-product of *how* the network learns to solve the task. As in the preceding section, the initial connectivity strength plays a role in the nature of this compression; Figure 3c shows how stronger initial connectivity leads to the formation of chaotic attractors that, while still compressed, are not as compressed as in the case of smaller initial connectivity strength (here both initializations are far from the *tabula rasa* regime).

Contemporaneously, a similar compression phenomenon was observed in convolutional neural networks trained on image recognition tasks [57] (Figure 3d); see also [59,62]. The work of Ref. [57] also discovered further very interesting aspects of the geometry of the representation relevant to the higher-dimensional output space, such as the compressed clusters lying on the vertices of a simplex and mathematically analyzed consequences of the resulting representations. We note that Ref. [63] highlighted a limitation of early studies of neural collapse, in that the phenomenon was examined for representations of training data and in cases may not be as robust for testing data.

Overall, we note that neural collapse adds additional structure beyond the low-rank representations that emerge in, for example, linear networks trained in the *tabula rasa* regime [42,43]. This is because the action of forming distinct localized clusters is highly nonlinear. This said, linearized analysis local to each separate cluster may still give insight into the underlying mechanisms [53,59], especially if this linearization can be justified in some limit as is done in Ref. [46]. See also Ref. [41] for an in-depth analysis of compression in two-layer neural networks.

Information bottleneck

The *information bottleneck* ideas and results by Tishby and colleagues preceded the papers above and played a highly influential role in showing how neural networks can compress task-irrelevant information overall [52,64]. The key quantity here is mutual information.

The authors demonstrate a very interesting phenomenon that can occur in deep neural networks: deep layers form representations that learn to maximize mutual information about task outputs, while minimizing overall information about network inputs. This is illustrated via the "learning curves" in Figure 3g, from Refs. [52,61]. For a given layer T, these curves track two quantities over the course of network learning. The first, $I(X, T)$, is the mutual information between representations in layer T and the network inputs, X. The second, $I(T, Y)$, is the corresponding information for the task outputs, Y. A key point is that, as learning evolves across epochs and $I(T, Y)$ continues to increase, eventually $I(X, T)$ begins to *decrease*. Thus, task-irrelevant mutual information is gradually compressed over the course of learning.

Remarkably, the authors also show that this compression can follow a precise optimization relationship, in which the mutual information about task inputs is minimized, subject to preserving mutual information about task outputs. This connects directly to earlier analytical work on the broader concept of an information bottleneck [65]. Moreover, the authors develop a mechanism by which this bottleneck may develop through the course of learning in neural networks. This stochasticity is inherent in incremental learning processes, in which network weights are incremented step by step based on the successive examples (or "batches" of examples) on which the network is trained. Because each is randomly selected, there is a deviation on each step from the true task gradient. Intriguingly, the authors show how this can, at least under certain assumptions, result in the same selective compression of task-irrelevant input information that leads to the information bottleneck found in their simulations and broader theory.

Robustness of exclusive representations

As in the results reviewed above showing that the emergence of rich vs. lazy representations depends (at least) on network initialization, the emergence of representations that compress task-irrelevant information is also far from automatic. Rather, it depends on many variables including network architecture, initialization scheme, loss function, and optimization procedure. In fact, for some tasks that lead to highly compressed representations in some settings, even changes to the loss function and details of gradient learning algorithm ("optimizer") can eliminate compression. As a striking example, the compression seen in Figure 3b vanishes when the optimizer is changed from RMSprop to "vanilla" stochastic gradient descent [53], and the network instead uses a lazy strategy to solve the task.

Given this possibility, when does neural collapse actually occur? The study [53] highlights three mechanisms that encourage neural collapse: the combination of loss functions that encourage scaling up outputs (such as categorical cross entropy) with saturating nonlinearities in the intermediate layers, large variability during the training process, and weight decay (in which additional terms are added to the network cost function to penalize large weights during training). The underlying theory generally uses a geometric decomposition of the learning dynamics into task-relevant and task-irrelevant directions [53,59,58]. In Figure 3e and f, we provide supporting evidence for this theory in the context of the deep convolutional neural network Resnet18 [66] trained on the image recognition task CIFAR-10 [67]. Here, we find that variability induced by dropout—-where neurons are randomly silenced, similar to a decrease in firing rate caused by a biological neuron's failure to spike—results in more compressed representations (see Figure 3e for stochastic gradient descent with momentum and Figure 3f for the RMSprop optimizer). This suggests that noise inherent in biological neural circuits may drive neural collapse. We use a simple measure of task-irrelevant compression, which is simply the average within-class distance divided by the average across-class distance of points in representation space. This collapse is measured on testing data, with mean squared error as a loss function (compare with Ref. [63] where categorical cross-entropy loss is used).

Similarly, not all networks, tasks, and training processes lead to information bottlenecks. For example Ref. [61] argues that this result can be limited to networks with double-sided saturating nonlinearities such as the hyperbolic tangent functions, again limiting the universality of the phenomenon (Figure 3h).

Titrating the representation: partial compression across time and across layers

A highly interesting aspect of the compression of task-irrelevant information in neural networks is that, even in fully trained networks, it is not necessarily an all-or-none process. Rather, several studies have shown that this compression can occur gradually across layers of deep networks, or gradually across time in recurrent neural networks. This gradation has potentially important implications in neuroscience: downstream brain areas with access to the neural network at different stages (layers or timesteps) would have access to different levels of stimulus information. This could be useful in driving different learning, memory, or behavioral systems downstream ([68] and John Maunsell, personal communication).

The graded nature of information compression is clearly evident in the information bottleneck formulation, where learning occurs in a two-phase process. In the first stage, intermediate layers first gain information about inputs and outputs (with input information dominating in the earlier layers and output information dominating in later layers). In the second phase, information about inputs is progressively lost (at a rate that is faster for later layers) [52,64] (Figure 3g). The authors of Ref. [53,59,62] also study collapse over layers or timesteps in the case of recurrent neural networks. They find that compression occurs progressively over the timesteps of the recurrent network (or over layers of a feedforward network). However, when trained on low-dimensional input data, the network first lifts the representation into a higher-dimensional space in the first few timesteps, reminiscent of the kernel machine approach, before potentially compressing the representation back down in later timesteps ([53], see also [69]).

Connections to experiments

Many studies have observed neural representations that are highly compressed when compared to nonlinear high-d representations. These include [3,9,48], which show that particular brain areas such as posterior parietal cortex are relatively compressed (see Figure 2h to j). In addition Ref. [3] shows that lateral intraparietal (LIP) neurons form a highly compressed representation that reflects the structure of task outputs and that adapts to new task structures. In contrast, the middle temporal areas have a lazy representation that reflects the structure of the inputs and does not adapt to changing task structure.

Less is known about whether certain brain areas/representations truly compress away information that is irrelevant to the task at hand, or at what temporal scale information is lost. To quote [3], "The exact nature of the role of LIP during learning, and whether changes in the … representations of LIP are stable or vary dynamically with the demands of the task, remain to be determined." While these questions still await more final answers, they can be approached by experimental paradigms that track learning over time. Indeed, the work of Stern and colleagues [60] both identified compression of representation dimension computationally and found evidence for this compression gradually emerging over days of task learning in wide-field activity patterns of mice. The degree to which the process of compressing task-irrelevant information—or "learning to forget"—can be tracked dynamically in the brain has important implications for our understanding of brain function, as well as remaining an exciting direction for further experimental studies.

Exclusive network representations in self-supervised tasks

Above, we have focused on supervised learning tasks, in which every input is paired with an intended output—for example, images and their labels. However, a vast and rapidly growing field studies neural networks trained on data without labels beyond that provided by the inputs themselves—a class of tasks known as semi-supervised or self-supervised (see e.g., Ref. [70–73]). Importantly, this setting likely corresponds to much of the learning that occurs in neurobiology as well, as in the real world, there is usually no explicit "teacher" constantly defining correct task outputs but rather sustained exposure to a richly informative sensory world.

Rather than attempting to cover the vast literature on self-supervised learning and representations, we ask the specific question that ties most immediately to the above: what are the consequences of rich and exclusive representations for self-supervised learning? We leave the possibility of lazy solutions to predictive learning as an interesting topic for further exploration.

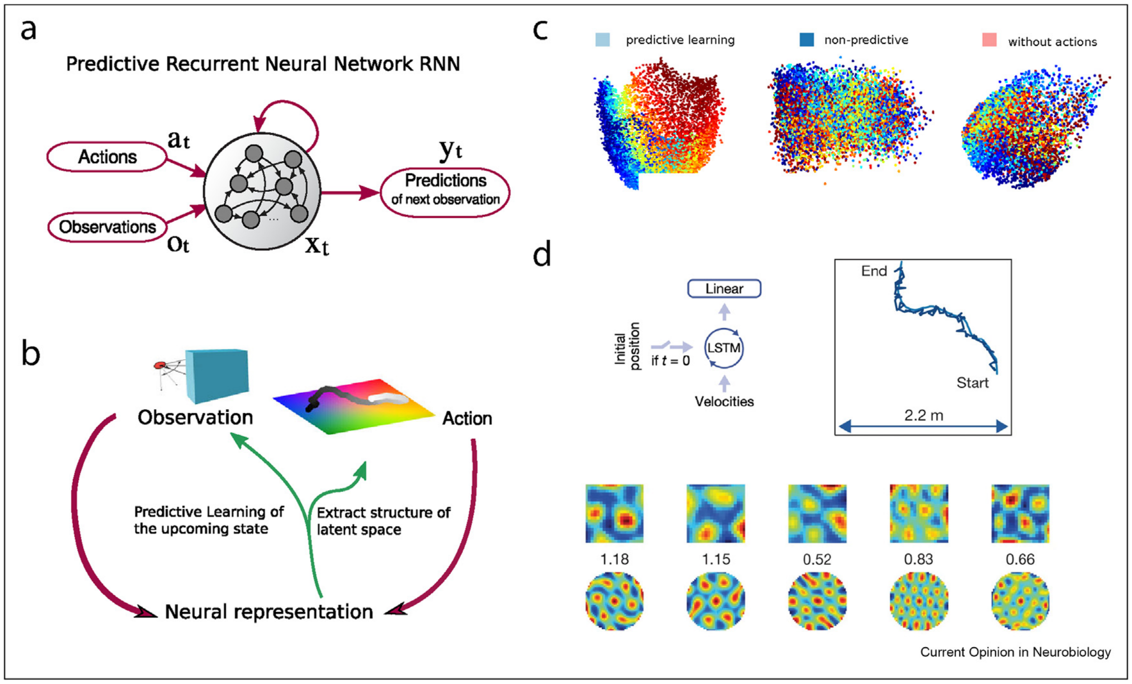
We focus on the prominent self-supervised setting of predictive learning. In predictive learning, inputs occur as a time series and networks are trained to predict their future values. Here, the target output of the network is a temporally shifted version of the input, which the network is trained to predict. Thus, task-relevant features are features that bear predictive power over future inputs. Conversely, task-irrelevant features do not influence future inputs.

Collapsed representations and the extraction of latent variables

A prominent example of a predictive learning task is navigation. In one recent study [74], a recurrent network was trained to predict future visual observations of an agent moving through an environment (Figure 4a and b). Here, the agent's actions were part of the input but not the output of the network, thus reflecting the agent's internal knowledge. In such a task the agent's field of view, and thus its visual observations, are completely determined by its state in the environment, the (x, y) location and orientation θ, so that (x, y, θ) are the underlying *latent variables* or *latent states*. This means that if the network could infer the current latent state (x, y, θ) from the visual inputs, it could in principle apply the agent's actions and anticipate future observations, through a downstream mapping from the latent space to observations. The work [74] showed that predictive learning extracts such latent states, through activity patterns that emerge in a trained recurrent neural network. Moreover, such latent states are less apparent in control networks trained to encode, but not predict, their inputs (Figure 4c).

In this sense, the predictive network learns a parsimonious model of the structure of its world, and rules for how to update the states defining that world based on its actions. In the process, detailed visual scene information is compressed. This results in relatively low-dimensional neural representations of the latent states. Recall that the task-relevant features here are those that enable the

Figure 4



Learning to represent information in the absence of labeled data: predictive learning in navigation tasks and emergent latent variables. (**a**)–(**c**) Adapted from Ref. [74]. (**a**) Illustration of the network architecture for a navigational predictive learning task. (**b**) Illustration of the task environment for a navigational predictive learning task. (**c**) The top three principal components for the representation of a neural network after training on the predictive task, colored by position and orientation, and compared with other versions of the task. Left: predictive learning. Center: non-predictive learning. Right: predictive learning without the action information. (**d**) Emergence of grid cells in a recurrent neural network trained to path integrate. Note that path integration is a form of prediction—velocity signals (actions) are used to predict future positions. Taken from Ref. [75] (reproduced with permission from Nature).

prediction of future inputs—that is, the latent states. Thus, the network has learned an internal representation that is compressed around these task-relevant features.

The above gives a concrete illustration of how the ideas of representing task structure and compressing task-irrelevant inputs can be applied to the setting of predictive learning: the sole distinction from the supervised cases above is that the concept of task-relevant structure has been replaced by the concept of latent-state structure [74]. Recently, biologically inspired Hebbian learning rules have also been derived that perform predictive learning [76,77], and the network behavior (and extracted latent variables) analyzed.

In theory, networks in the lazy regime will not form structured representations such as these. As far as we are aware, whether or not networks in the lazy regime are able to satisfactorily perform predictive learning remains an open avenue for future work.

Connections to neuroscience

Several influential lines of work propose that predictive learning is a major driver of neural representations across the brain [78–80]. This said, the modeling approaches proposed for different brain areas can differ. Models focusing on the visual stream (i.e. sensory prediction) have historically been decoupled from those focusing on hippocampal dynamics (i.e. memory-based prediction). Moreover, the former models have placed emphasis on spatial prediction (e.g. completing missing elements of an image) and the latter on temporal prediction (similar to that discussed above). In the first instance, predictive models have been found to extract representations that reproduce visual receptive fields [81] and other properties of how sensory systems encode information [76,82]. In the second, the majority of research has centered on navigation, demonstrating that during predictive learning neurons begin to tile location and orientation in their activations [74]. This is similar to place cells and head direction cells in navigation-related brain circuits [83].

The case of grid cells merits special consideration. In Ref. [75], the authors trained a network to directly predict the next (x, y) location based on the current location, orientation, and the action the agent takes (a calculation known as path integration) (Figure 4d). The authors found that many of the units in the trained network functioned as grid cells, thus adopting a well-known encoding of spatial latent variables. While a theoretical accounting of this phenomenon was provided by Ref. [84], the sensitivity of grid cells' emergence to choice network architectures, training rules, and allied hyperparameters is a topic of ongoing research [75,85,86]. As one example, the authors of Ref. [75] note dropout as an important mechanism in the appearance of grid cell representations, perhaps connecting to work on the principles of neural collapse reviewed above.

Discussion

The prevalence of task-trained neural networks as models of the brain is exploding. This makes understanding the robustness and universality of neural network behavior essential, so that we can properly contextualize the insights these behaviors may provide for neural circuits in the brain. In this spirit, we began our review with a fundamental but remarkably subtle question: when do such networks learn to encode tasks in an observable way? First, we introduced multiple regimes to determine when neural networks learn task features in their internal representations. Then, we investigated the allied phenomenon of forming compressed representations that isolate information that is relevant to the task at hand. Finally, we demonstrated how these principles arise in self-supervised learning.

The emergence of task-relevant features in neural representations is not a given

While both feature learning and compressed/exclusive representations occur widely, they do not occur all the time. This is a major takeaway from our review: the details of learning algorithms and network structure matter. It may not be sufficient to train a neural network and examine the learnt solutions; rather, it is necessary to verify, and illuminating to explore, the robustness of the findings about network representations with respect to a number of factors, including initialization and optimization schemes. One extreme example is that neural networks can sometimes learn to solve tasks by merely updating output weights during training, or by updating weights in a manner that does not truly lead to learning task-driven features (such as with NTK initialization). At another extreme, they may begin with no structure and learn task-driven features alone (*tabula rasa*), or begin with task-irrelevant structure that is removed through training (compression/neural collapse). Intermediate outcomes are also possible.

As such, training a single network as a model of a brain circuit is unlikely to bear definitive results. Rather, we will need to evaluate the spectrum of responses that neural networks can exhibit, or clearly justify specific choices of their initialization, architecture, and learning rules. Theory can help guide the way here. As we reviewed, some of the mechanisms that appear to encourage task-structured representations include weight decay, "noisy" optimization processes such as RMSprop and ADAM, saturating nonlinearities, and added sources of network noise such as dropout. This said, much remains to be understood about the role of these and other network and learning mechanisms.

Reinforcement learning, and the revenge of the "hand-built" model

While we have focused this review on the behavior of trained neural networks, there are many other domains in which the distinction between lazy and rich solutions is highly relevant. One example is networks trained via reinforcement learning algorithms. It is still very much an open challenge to delineate the two regimes in this domain and to work out the functional consequences of the different approaches.

Another important example is networks where many components are not trained, but "hand-built" with a given functionality in mind. In this approach, typically a theorist has in mind a desired network function and uses geometrical and mathematical reasoning to build networks that fulfill it. In some instances, only the output weights are trained, and in others, there is no need for training at all. An example of the latter is attractor networks hand-built to integrate velocity signals in order to track position, such as models of head-direction cells [83] and grid cells [87]. Indeed, linking network models of brain circuits provided by gradient descent training procedures and those that could be, at least in principle, built by hand, provides a fruitful path toward interpreting and understanding how the trained networks actually work (cf. [88]). While hand-built models tend to resemble networks in the rich regime, some models, such as reservoir computers, are designed to be in the lazy regime.

Predictive learning in language models and beyond

We end our review with a discussion on the burgeoning area of self-supervised predictive learning. Prediction forces a learning agent to learn how operations and actions influence the world. Such an understanding is likely best supported by an efficient representation that reflects the relatively low-dimensional structure of the latent variables, though this remains to be proved. Here, we discussed likely connections to feature learning and information compression. However, theory describing the structure (or lack thereof) extracted by predictive learning, especially under the wide possibilities of different initializations, optimization schemes, and other hyperparameters, is still in its infancy.

Language models are a particularly topical example of the ability of predictive models to extract latent space information. While we make no attempt to review the tremendous advances made in natural language processing (NLP) over the past decade, we note that the underlying models have employed multiple prediction-based techniques to extract language structure. For instance, the famous project word2vec [89] reveals that neural networks trained to predict omitted words acquire a representation of these words that forms a latent space map corresponding to word meanings. With this representation, word-based manipulations appear meaningful (e.g. King - Man + Woman = Queen). Recent theoretical work has enhanced our understanding of such vector-symbolic operations [90–92]. The efficacy of allied models [93] in extracting the fundamental structure of languages has led to breakthroughs such as ChatGPT and crosslingual translation between any two languages [94–96].

This said, while predictive training has been a cornerstone in developing language models, experimental tests of whether the underlying representations also appear in biological brains have been limited by the difficulty of conducting language-based studies in nonhuman animals. Nevertheless, recent research indicates that human-level language comprehension involves predictive processing [97,98]. Finally, we note that implicit in the above is our speculation that the emergence of latent variable structure in the representation of language is driven, at least in part, by the same factors reviewed above that promote the compression of task-irrelevant information. Verifying or rejecting this speculation is an intriguing target for future modeling and theoretical work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

We thank the reviewers for their careful and thoughtful suggestions which have substantially improved this manuscript. MF thanks Jacob Zavatone-Veth and Blake Bordelon for their guidance regarding technical aspects of rich versus lazy learning. MF also thanks Cengiz Pehlevan for guidance in conducting experiments for probing neural collapse. E. S-B. gratefully acknowledges support of NIH Grants UF1NS126485 and RF1DA055669 and of NSF NCS-FO Grant 2024364. All authors gratefully thank the Swartz Foundation and its Swartz Program in Theoretical Neuroscience at Harvard (MF) and Swartz Center for Theoretical Neuroscience at the University of Washington for their support.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Rigotti Mattia, Barak Omri, Warden Melissa R, Wang Xiao-Jing, Daw Nathaniel D, Miller Earl K, Fusi Stefano: **The importance of mixed selectivity in complex cognitive tasks**. *Nature* May 2013, **497**:585–590.
2. Fusi Stefano, Miller Earl K, Rigotti Mattia: **Why neurons mix: high dimensionality for higher cognition**. *Curr Opin Neurobiol* April 2016, **37**:66–74.
3. Freedman David J, Assad John A: **Experience-dependent representation of visual categories in parietal cortex**. *Nature* September 2006, **443**:85–88.
4. Barak Omri: **Recurrent neural networks as versatile tools of neuroscience research**. *Curr Opin Neurobiol* 2017, **46**:1–6.
5. Rumelhart David E, McClelland James L: *Parallel distributed processing: explorations in the microstructure of cognition: foundations*. MIT Press; 1987.

6. Yang Guangyu Robert, Wang Xiao-Jing: **Artificial neural networks for neuroscientists: a primer**. *Neuron* 2020, **107**: 1048–1070.

7. Mante Valerio, Sussillo David, Shenoy Krishna V, Newsome William T: **Context-dependent computation by recurrent dynamics in prefrontal cortex**. *Nature* November 2013, **503**:78–84.

8. Yamins Daniel LK, Ha Hong, Cadieu Charles F, Solomon Ethan A, Seibert Darren, DiCarlo James J: **Performance-optimized hierarchical models predict neural responses in higher visual cortex**. *Proc Natl Acad Sci USA* 2014, **111**:8619–8624.

9. ** Timo Flesch, Juechems Keno, Dumbalska Tsvetomira, Saxe Andrew, Summerfield Christopher: **Orthogonal representations for robust context-dependent task performance in brains and neural networks**. *Neuron* April 2022, **110**: 1258–1270.e11.

Investigation of whether the brain performs lazy or rich learning of task structure.

10. Arthur Jacot, Gabriel Franck, Clement Hongler: **Neural tangent kernel: convergence and generalization in neural networks**. *Adv Neural Inf Process Syst* 2018, **10**.

11. ** Yang Greg, Edward J, Hu: **Tensor programs iv: feature learning in infinite-width neural networks**. In *International conference on machine learning*; 2021.

In-depth study of the (lack of) feature learning that occurs in the neural tangent kernel regime and beyond. This study parameterizes neural network initializations and shows how the feature learning behavior changes through a large range of these parameterizations.

12. Zavatone-Veth Jacob A, Canatar Abdulkadir, Ruben Benjamin S, Pehlevan Cengiz: **Asymptotics of representation learning in finite Bayesian neural networks**. *J Stat Mech Theor Exp* November 2022, **2022**:114008.

13. Bordelon Blake, Pehlevan Cengiz: **The influence of learning rule on representation dynamics in wide neural networks**. In *The eleventh international conference on learning representations*; 2023.

14. Zhao Haitao, Lai Zhihui, Leung Henry, Zhang Xianyi: *Feature learning and understanding: algorithms and applications*. Springer Cham; 2020.

15. Chung SueYeon, Abbott LF: **Neural population geometry: an approach for understanding biological and artificial neural networks**. *Curr Opin Neurobiol* 2021, **70**:137–144.

16. Steinwart Ingo, Christmann Andreas: *Support vector machines*. 1st ed. New York: Information science and statistics. Springer; 2008.

17. Ali Rahimi, Recht Benjamin: **Uniform approximation of functions with random bases**. In *2008 46th annual allerton conference on communication, control, and computing*; 2008:555–561.

18. Schölkopf Bernhard, Smola Alexander J: *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press; 2002.

19. Maass Wolfgang, Natschläger Thomas, Henry Markram: **Real-time computing without stable states: a new framework for neural computation based on perturbations**. *Neural Comput* November 2002, **14**:2531–2560.

20. Jaeger Herbert: **The "echo state" approach to analysing and training recurrent neural networks-with an erratum note**. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report* January 2001, **148**:1–47.

21. Mezić Igor: **Koopman operator, geometry, and learning of dynamical systems**. *Not AMS* August 2021, **68**:1.

22. Williams Christopher: **Computing with infinite networks**. In Mozer MC, Jordan M, Petsche T. *Advances in neural information processing systems*, **9**. MIT Press; 1996.

23. Neal Radford M: *Priors for infinite networks*. New York, NY: Springer New York; 1996:29–53.

24. Lee Jaehoon, Sohl-dickstein Jascha, Pennington Jeffrey, Novak Roman, Schoenholz Sam, Bahri Yasaman: **Deep neural networks as Gaussian processes**. In *International conference on learning representations*; 2018.

25. Matthews Alexander G de G, Hron Jiri, Rowland Mark, Turner Richard E, Ghahramani Zoubin: **Gaussian process behaviour in wide deep neural networks**. In *International conference on learning representations*; 2018.

26. Chase Steven, Schwartz Andrew, Maass Wolfgang, Legenstein Robert: **Functional network reorganization in motor cortex can be explained by reward-modulated Hebbian learning**. In Bengio Y, Schuurmans D, Lafferty J, Williams C, Culotta A. *Advances in neural information processing systems*, **22**. Curran Associates, Inc.; 2009:1105–1113.

27. Litwin-Kumar Ashok, Harris Kameron Decker, Axel Richard, Sompolinsky Haim, Abbott LF: **Optimal degrees of synaptic connectivity**. *Neuron* March 2017, **93**:1153–1164.e7.

28. Lanore Frederic, Cayco-Gajic N Alex, Gurnani Harsha, Coyle Diccon, Silver R Angus: **Cerebellar granule cell axons support high-dimensional representations**. *Nat Neurosci* June 2021, **24**:1142–1150.

29. Cayco-Gajic N Alex, Clopath Claudia, Silver R Angus: **Sparse synaptic connectivity is required for decorrelation and pattern separation in feedforward networks**. *Nat Commun* October 2017, **8**:1–11.

30. ** Schuessler Friedrich, Mastrogiuseppe Francesca, Dubreuil Alexis, Ostojic Srdjan, Barak Omri: **The interplay between randomness and structure during learning in rnns**. In Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H. *Advances in neural information processing systems*, **33**. Curran Associates, Inc.; 2020:13352–13362.

Study of the dependence of the learning dynamics of a recurrent neural network on initial recurrent connectivity coupling strength. Includes a mathematical analysis of training on simple tasks that predicts changes in weight structure through learning.

31. Alemohammad Sina, Wang Zichao, Balestriero Randall, Baraniuk Richard: **The recurrent neural tangent kernel**. In *International conference on learning representations*; 2021.

32. Yang Greg, Littwin Etai: **Tensor programs iib: architectural universality of neural tangent kernel training dynamics**. In Proceedings of the 38th international Conference on machine learning*, volume 139 of* Proceedings of machine learning research. Edited by Meila Marina, Zhang Tong; Jul 2021: 11762–11772. PMLR, 18–24.

33. Schiller Ulf D, Steil Jochen J: **Analyzing the weight dynamics of recurrent learning algorithms**. *Neurocomputing* 2005, **63**: 5–23.

34. ** Chizat Lénaïc, Oyallon Edouard, Bach Francis: **On Lazy Training in Differentiable Programming**. *Advances in Neural Information Processing Systems* 2019, **32**.

Early observations of and mathematical arguments supporting the existence of the phenomenon of lazy training. Here it is first observed that lazy training is related to the scaling of the initial weights.

35. Saxe Andrew, Sodhani Shagun, Jay Lewallen Sam: **The neural race reduction: dynamics of abstraction in gated networks**. In *Proceedings of the 39th international Conference on machine learning, volume 162 of Proceedings of machine learning research*. Edited by Chaudhuri Kamalika, Jegelka Stefanie, Song Le, Szepesvari Csaba, Niu Gang, Sivan Sabato, PMLR; Jul 2022:19287–19309. 17–23.

36. * Atanasov Alexander, Bordelon Blake, Pehlevan Cengiz: **Neural networks as kernel learners: the silent alignment effect**. In *International conference on learning representations*; 2022.

Study of neural network learning that proposes that the evolution of the neural tangent kernel occurs in two phases. This assumption holds under *tabula rasa* initialization, but is shown to also be approximately true beyond the *tabula rasa* regime.

37. Dyer Ethan, Gur-Ari Guy: **Asymptotics of wide networks from feynman diagrams**. In *International conference on learning representations*; 2020.

38. Li Qianyi, Sompolinsky Haim: **Statistical mechanics of deep linear neural networks: the backpropagating kernel renormalization**. *Phys Rev X* Sep 2021, **11**:31059.

39. Zavatone-Veth Jacob A, Tong William L, Pehlevan Cengiz: **Contrasting random and learned features in deep bayesian linear regression**. *Phys Rev E* Jun 2022, **105**:64118.

40. Roberts Daniel A, Yaida Sho, Hanin Boris: *The principles of deep learning theory*. Cambridge University Press; 2022. https://deeplearningtheory.com.

41. Jonas Paccolat, Petrini Leonardo, Geiger Mario, Kevin Tyloo, Wyart Matthieu: **Geometric compression of invariant manifolds in neural networks**. *J Stat Mech Theor Exp* Apr 2021, **2021**:44001.

42. Saxe Andrew M, McClelland James L, Ganguli Surya: **Exact solutions to the nonlinear dynamics of learning in deep linear neural networks**. In *International Conference on Learning Representations 2014*; February 2014.

43. Saxe Andrew M, McClelland James L, Ganguli Surya: **A mathematical theory of semantic development in deep neural networks**. *Proc Natl Acad Sci USA* June 2019, **116**: 11537−11546.

44. Shi Jianghong, Todd Shea-Brown Eric, Buice Michael A: **Learning dynamics of deep linear networks with multiple pathways**. In *Advances in neural information processing systems*. Edited by Oh Alice H, Agarwal Alekh, Belgrave Danielle, Cho Kyunghyun; 2022.

45. * Braun Lukas, Juliette Dominé Clémentine Carla, Fitzgeral James E, Saxe Andrew: **Exact learning dynamics of deep linear networks with prior knowledge**. In *Advances in neural information processing systems*. Edited by Oh Alice H, Agarwal Alekh, Belgrave Danielle, Cho Kyunghyun; 2022.

Study of the learning dynamics for deep linear neural networks, taking into account various possible initializations. This theory sheds light on the transition of networks from rich to lazy learning, in which internal representations do and do not learn task structure.

46. Mastrogiuseppe Francesca, Hiratani Naoki, Latham Peter: **Evolution of neural activity in circuits bridging sensory and abstract knowledge**. *Elife* March 2023, **12**, e79908.

47. Ito Takuya, John D, Murray: **Multitask representations in the human cortex transform along a sensory-to-motor hierarchy**. *Nat Neurosci* February 2023, **26**:306−315.

48. Jeffrey Johnston W, Fusi Stefano: **Abstract representations emerge naturally in neural networks trained to perform multiple tasks**. *Nat Commun* February 2023, **14**:1040.

49. Dubreuil Alexis, Valente Adrian, Beiran Manuel, Mastrogiuseppe Francesca, Ostojic Srdjan: **The role of population structure in computations through neural dynamics**. *Nat Neurosci* June 2022, **25**:783−794.

50. Hirokawa Junya, Vaughan Alexander, Masset Paul, Ott Torben, Kepecs Adam: **Frontal cortex neuron types categorically encode single decision variables**. *Nature* December 2019, **576**: 446−451.

51. Raposo David, Kaufman Matthew T, Churchland Anne K: **A category-free neural population supports evolving demands during decision-making**. *Nat Neurosci* December 2014, **17**:1784−1792.

52. Tishby Naftali: *The information bottleneck theory of deep neural networks*. Bulletin of the American Physical Society; 2018.

53. ** Farrell Matthew, Recanatesi Stefano, Moore Timothy, Lajoie Guillaume, Shea-Brown Eric: **Gradient-based learning drives robust representations in recurrent neural networks by balancing compression and expansion**. *Nat Mach Intell* June 2022, **4**:1−10.

Study of the change in representation induced by training a recurrent neural network on a classification task. Depending on initialization and task, the network expands or compresses its representation, or a combination of both; factors that encourage compression are identified empirically and through some analytical approximations of learning dynamics.

54. Canatar Abdulkadir, Bordelon Blake, Pehlevan Cengiz: **Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks**. *Nat Commun* May 2021, **12**:2914.

55. Bernardi Silvia, Benna Marcus K, Rigotti Mattia, Munuera Jérôme, Fusi Stefano, Daniel Salzman C: **The geometry of abstraction in the hippocampus and prefrontal cortex**. *Cell* 2020, **183**: 954−967.e21.

56. Farrell Matthew: *Revealing structure in trained neural networks through dimensionality-based methods*. University of Washington ProQuest Dissertations Publishing, 113, 1-116; 2020.

57. ** Papyan Vardan, Han XY, Donoho David L: **Prevalence of neural collapse during the terminal phase of deep learning training**. *Proc Natl Acad Sci USA* October 2020, **117**:24652−24663.

Empirical study of geometric compression that occurs in the penultimate layers of image classification networks, with rich mathematical analysis of consequences. This work introduced a formal definition of compression, called neural collapse, which has inspired many follow-up studies.

58. Han XY, Papyan Vardan, Donoho David L: **Neural collapse under MSE loss: proximity to and dynamics on the central path**. In *International conference on learning representations*; 2022.

59. Recanatesi Stefano, Farrell Matthew, Advani Madhu, Moore Timothy, Lajoie Guillaume, Shea-Brown Eric: **Dimensionality compression and expansion in deep neural networks**. *arXiv:1906.00443 [cs, stat]* October 2019.

60. Stern M, *et al.*: *In the footsteps of learning: changes in network dynamics and dimensionality with task acquisition*. COSYNE Conference Abstract; 2020.

61. Michael Saxe Andrew, Bansal Yamini, Joel Dapello, Advani Madhu, Kolchinsky Artemy, Daniel Tracey Brendan, Cox David Daniel: **On the information bottleneck theory of deep learning**. In *International conference on learning representations*; 2018.

62. Ansuini Alessio, Laio Alessandro, Macke Jakob H, Zoccolan Davide: **Intrinsic dimension of data representations in deep neural networks**. *NIPS* 2019, **11**.

63. Hui Like, Belkin Mikhail, Nakkiran Preetum: **Limitations of neural collapse for understanding generalization in deep learning**. *arXiv:2202.08384v1 [cs.LG]* 2022.

64. Shwartz-Ziv Ravid, Tishby Naftali: **Opening the black box of deep neural networks via information**. *arXiv:1703.00810v3 [cs.LG]* March 2017.

65. Tishby Naftali, Pereira Cicero, Bialek William: **The information bottleneck method**. *Proceedings of the 37th Allerton Conference on Communication, Control and Computation* 2001, **49**:7.

66. He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian: **Deep residual learning for image recognition**. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)*; 2016: 770−778.

67. Krizhevsky Alex: *Learning multiple layers of features from tiny images*. Technical report, University of Toronto; 2009.

68. Stephenson Cory, Padhy Suchismita, Ganesh Abhinav, Yue Hui, Tang Hanlin, Chung SueYeon: **On the geometry of generalization and memorization in deep neural networks**. In *International conference on learning representations*; 2021.

69. Keup Christian, Tobias Kühn, Dahmen David, Helias Moritz: **Transient chaotic dimensionality expansion by recurrent networks**. *Phys Rev X* Jun 2021, **11**:21064.

70. Jing Longlong, Tian Yingli: **Self-supervised visual feature learning with deep neural networks: a survey**. *IEEE Trans Pattern Anal Mach Intell* 2020, **43**:4037−4058.

71. Aaron van den Oord, Li Yazhe, Oriol Vinyals: **Representation learning with contrastive predictive coding**. *arXiv:1807.03748 [cs, stat]* January 2019.

72. Hinton Geoffrey, Sejnowski Terrence J: *Unsupervised learning: Foundations of neural computation*. The MIT Press; May 1999.

73. Balestriero Randall, Ibrahim Mark, Vlad Sobal, Morcos Ari, Shekhar Shashank, Goldstein Tom, Bordes Florian, Bardes Adrien, Mialon Gregoire, Tian Yuandong, Schwarzschild Avi, Wilson Andrew Gordon, Jonas Geiping, Garrido Quentin, Fernandez Pierre, Bar Amir, Pirsiavash Hamed,

LeCun Yann, Goldblum Micah: **A cookbook of self-supervised learning**. *arXiv:2304.12210 [cs]* June 2023.

74. Recanatesi Stefano, Farrell Matthew, Lajoie Guillaume, Deneve Sophie, Rigotti Mattia, Shea-Brown Eric: **Predictive learning as a network mechanism for extracting low-dimensional latent space representations**. *Nat Commun* March 2021, **12**:1417.
* *

Study of the representations that emerge in recurrent neural networks trained to predict the future of their inputs. In this study, an agent moves randomly through an environment, and the network is trained to predict the agent's next observation. This results in a spatial map of the agent's location emerging in the network representation over the course of learning.

75. Banino Andrea, Barry Caswell, Uria Benigno, Blundell Charles, Lillicrap Timothy, Mirowski Piotr, Alexander Pritzel, Chadwick Martin J, Degris Thomas, Joseph Modayil, Wayne Greg, Hubert Soyer, Viola Fabio, Zhang Brian, Ross Goroshin, Rabinowitz Neil, Pascanu Razvan, Beattie Charlie, Petersen Stig, Sadik Amir, Gaffney Stephen, King Helen, Kavukcuoglu Koray, Hassabis Demis, Hadsell Raia, Kumaran Dharshan: **Vector-based navigation using grid-like representations in artificial agents**. *Nature* 2018, **557**:429–433.

76. Lipshutz David, Windolf Charles, Golkar Siavash, Chklovskii Dmitri: **A biologically plausible neural network for slow feature analysis**. *Adv Neural Inf Process Syst* 2020, **33**: 14986–14996.

77. Halvagal Manu Srinath, Zenke Friedemann: **The combination of hebbian and predictive plasticity learns invariant object representations in deep sensory networks**. *bioRxiv* 2023.

78. Teufel Christoph, Fletcher Paul C: **Forms of prediction in the nervous system**. *Nat Rev Neurosci* 2020, **21**:231–242.

79. Eichenbaum Howard, Norbert J Fortin: **The neurobiology of memory based predictions**. *Phil Trans Biol Sci* 2009, **364**: 1183–1191.

80. Rao Rajesh PN: **A sensory-motor theory of the neocortex based on active predictive coding**. *bioRxiv* 2022–12:2022.

81. Lotter William, Gabriel Kreiman, Cox David: **Deep predictive coding networks for video prediction and unsupervised learning**. *International Conference on Learning Representations* 2017.

82. Huang Yanping, Rao Rajesh PN: **Predictive coding**. *Wiley Interdisciplinary Reviews: Cognit Sci* 2011, **2**:580–593.

83. Khona Mikail, Ila R: **Fiete. Attractor and integrator networks in the brain**. *Nat Rev Neurosci* December 2022, **23**:744–766.

84. Ben Sorscher, Mel Gabriel C, Ocko Samuel A, Giocomo Lisa M, Ganguli Surya: **A unified theory for the computational and mechanistic origins of grid cells**. *Neuron* January 2023, **111**: 121–137.e13.

85. Schaeffer Rylan, Khona Mikail, Fiete Ila R: **No free lunch from deep learning in neuroscience: a case study through models of the entorhinal-hippocampal circuit**. In *Advances in neural information processing systems*. Edited by Oh Alice H, Agarwal Alekh, Belgrave Danielle, Cho Kyunghyun; 2022.

86. Ben Sorscher, Mel Gabriel C, Nayebi Aran, Giocomo Lisa, Yamins Daniel, Ganguli Surya: **When and why grid cells appear or not in trained path integrators**. *bioRxiv* 2022.

87. Burak Yoram, Fiete Ila R.: **Accurate path integration in continuous attractor network models of grid cells**. *PLoS Comput Biol* February 2009, **5**, e1000291.

88. Vyas D Sussillo S, Golub MD, Shenoy KV: **Computation through neural population dynamics**. *Annu Rev Neurosci* 2020, **43**:249–275.

89. Mikolov Tomas, Chen Kai, Corrado Greg, Dean Jeffrey: **Efficient estimation of word representations in vector space**. *arXiv: 1301.3781 [cs.CL]* 2013.

90. Hiratani Naoki, Sompolinsky Haim: **Optimal quadratic binding for relational reasoning in vector symbolic neural architectures**. *Neural Comput* January 2023, **35**:105–155.

91. Frady Edward Paxon, Kleyko Denis, Sommer Friedrich T: **Variable binding for sparse distributed representations: theory and applications**. *IEEE Transact Neural Networks Learn Syst* 2021:1–14.

92. Maheswaranathan Niru, Williams Alex H, Golub Matthew D, Ganguli Surya, Sussillo David: **Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics**. *Adv Neural Inf Process Syst* December 2019, **32**: 15696–15705.

93. Mars Mourad: **From word embeddings to pre-trained language models: a state-of-the-art walkthrough**. *Appl Sci* 2022, **12**: 8805.

94. Brown Tom, Mann Benjamin, Ryder Nick, Subbiah Melanie, Kaplan Jared D, Dhariwal Prafulla, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda, *et al.*: **Language models are few-shot learners**. *Adv Neural Inf Process Syst* 2020, **33**:1877–1901.

95. Conneau Alexis, Khandelwal Kartikay, Goyal Naman, Chaudhary Vishrav, Guillaume Wenzek, Guzmán Francisco, Grave Edouard, Ott Myle, Zettlemoyer Luke, Stoyanov Veselin: **Unsupervised cross-lingual representation learning at scale**. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. Association for Computational Linguistics; July 2020:8440–8451. Online.

96. Liu Pengfei, Yuan Weizhe, Fu Jinlan, Jiang Zhengbao, Hayashi Hiroaki, Graham Neubig: **Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing**. *ACM Comput Surv* 2023, **55**: 1–35.

97. Heilbron Micha, Armeni Kristijan, Schoffelen Jan-Mathijs, Peter Hagoort, De Lange Floris P: **A hierarchy of linguistic predictions during natural language comprehension**. *Proc Natl Acad Sci USA* 2022, **119**, e2201968119.

98. Nieuwland Mante S: **Do “early”’ brain responses reveal word form prediction during language comprehension? a critical review**. *Neurosci Biobehav Rev* 2019, **96**:367–400.