frontiers | Frontiers in Neural Circuits

TYPE Perspective
PUBLISHED 29 August 2025
DOI 10.3389/fncir.2025.1618351

OPEN ACCESS

EDITED BY
Nicoletta Berardi,
University of Florence, Italy

REVIEWED BY
Alexander van Meegen,
Swiss Federal Institute of Technology Lausanne, Switzerland

*CORRESPONDENCE
Jacob A. Zavatone-Veth
✉ jzavatoneveth@fas.harvard.edu
Blake Bordelon
✉ blake_bordelon@g.harvard.edu
Cengiz Pehlevan
✉ cpehlevan@seas.harvard.edu

RECEIVED 25 April 2025
ACCEPTED 11 August 2025
PUBLISHED 29 August 2025

CITATION
Zavatone-Veth JA, Bordelon B and Pehlevan C (2025) Summary statistics of learning link changing neural representations to behavior. *Front. Neural Circuits* 19:1618351. doi: 10.3389/fncir.2025.1618351

Summary statistics of learning link changing neural representations to behavior

Jacob A. Zavatone-Veth[1,2]*, Blake Bordelon[1,3]* and Cengiz Pehlevan[1,3,4]*

[1]Center for Brain Science, Harvard University, Cambridge, MA, United States, [2]Society of Fellows, Harvard University, Cambridge, MA, United States, [3]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, United States, [4]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA, United States

How can we make sense of large-scale recordings of neural activity across learning? Theories of neural network learning with their origins in statistical physics offer a potential answer: for a given task, there are often a small set of summary statistics that are sufficient to predict performance as the network learns. Here, we review recent advances in how summary statistics can be used to build theoretical understanding of neural network learning. We then argue for how this perspective can inform the analysis of neural data, enabling better understanding of learning in biological and artificial neural networks.

KEYWORDS

neural networks, learning, statistical physics, representation learning, summary statistics, representational similarity analysis

1 Introduction

Experience reshapes neural population activity, molding an animal's representations of the world as it learns to perform new tasks. Thanks to advances in experimental technologies, it is just now becoming possible to measure changes in the activity of large neural populations across the course of learning (Masset et al., 2022; Fink et al., 2025; Kriegeskorte and Wei, 2021; Steinmetz et al., 2021; Zhong et al., 2025; Sun et al., 2025; Vaidya et al., 2025). However, with this new capability comes the challenge of identifying which features of high-dimensional activity patterns are meaningful for understanding learning. While analyses of representations have begun how to elucidate how learning reshapes the structure of activity, it is not in general clear whether these measurements are sufficient to understand how representational changes relate to behavior (Krakauer et al., 2017; Sucholutsky et al., 2024; Kriegeskorte et al., 2008; Kriegeskorte and Wei, 2021).

In this Perspective, we propose that the principled identification of **summary statistics of learning** offers a possible path forward. This framework is grounded in theories of the statistical physics of learning in neural networks, which show that low-dimensional summary statistics are often sufficient to predict task performance over the course of learning (Watkin et al., 1993; Engel and van den Broeck, 2001; Zdeborová and Krzakala, 2016). We argue that thinking systematically about summary statistics gives new insight into what existing approaches of quantifying neural representations reveal about learning, and allows identification of what additional measurements would be required to constrain models of plasticity. We emphasize that the goal of this Perspective is not to advocate for the use of a particular set of summary statistics, but rather to explain the general philosophy of this approach to understanding learning in high dimensions.

2 What is a summary statistic?

We posit that summary statistics of learning must satisfy two minimal desiderata:

1. **They must be low-dimensional.** That is, their dimension is low relative to the number of neurons in the network of interest. Indeed, most summary statistics we will encounter are determined by averages over the population of neurons.
2. **They must be sufficient to predict behavior across learning.** From a theoretical standpoint, there should exist a closed set of equations describing the evolution of the summary statistics that predict the network's performance.

As we will illustrate with concrete examples in Section 3, summary statistics satisfying these two desiderata are often highly interpretable thanks to their clear relationship to the network architecture and learning task. However, the summary statistics relevant for predicting performance may not be sufficient to predict all statistical properties of population activity. We will elaborate on this issue, and the resulting limitations of descriptions based on summary statistics alone, in Section 4.

Our use of the term "summary statistics" follows work by Ben Arous et al. (2022, 2023). In the literature on the statistical physics of learning, the quantities that we refer to as summary statistics are often termed "order parameters" (Mézard et al., 1987; Watkin et al., 1993; Engel and van den Broeck, 2001; Zdeborová and Krzakala, 2016). We prefer to use the former, more general term as it better captures the goal of these reduced descriptions in a neuroscientific context: we aim to summarize the features of neural activity relevant for learning.

3 Summary statistics in theories of neural network learning

We now review how summary statistics emerge naturally in theoretical analyses of neural network learning. Out of many theoretical results, we focus on two example settings: online learning from high-dimensional data in shallow networks, and batch learning in wide and deep networks (Ben Arous et al., 2023; Goldt et al., 2019; Saad and Solla, 1995; Cui et al., 2023; Zavatone-Veth and Pehlevan, 2021; Bordelon and Pehlevan, 2023b; Zavatone-Veth et al., 2022b; Saxe et al., 2013; Bordelon et al., 2025; Arnaboldi et al., 2023; van Meegen and Sompolinsky, 2025; Watkin et al., 1993; Engel and van den Broeck, 2001; Zdeborová and Krzakala, 2016). These model problems illustrate how relevant summary statistics may be identified given a task, network architecture, and learning rule.

3.1 Online learning in shallow neural networks with high dimensional data

Classical models of online gradient descent learning in high dimensions can be often be summarized with simple summary statistics (Watkin et al., 1993; Engel and van den Broeck, 2001; Ben Arous et al., 2022; Arnaboldi et al., 2023; Goldt et al., 2019, 2020; Biehl and Schwarze, 1995; Saad and Solla, 1995). In this section, we discuss how the generalization performance of perceptrons and shallow (two-layer) neural networks trained on large quantities of high dimensional data can be summarized by simple weight alignment measures. Most simply, the perceptron model $f(\mathbf{x}) = \sigma\left(\frac{1}{\sqrt{D}}\mathbf{w}\cdot\mathbf{x}\right)$ seeks to learn a weight vector $\mathbf{w} \in \mathbb{R}^D$ which correctly classifies a finite set of randomly sampled training input-output pairs $(\mathbf{x}_\mu, y_\mu)$. If the inputs are random, $\mathbf{x}_\mu \sim \mathcal{N}(0, \mathbf{I}_D)$, and the targets $y_\mu = y(\mathbf{x}_\mu)$ are generated by a **teacher network** $y(\mathbf{x}) = \sigma\left(\frac{1}{\sqrt{D}}\mathbf{w}_\star\cdot\mathbf{x}\right)$, then the generalization performance (performance of the model on new *unseen data*, $\mathbb{E}_\mathbf{x}[(f(\mathbf{x}) - y(\mathbf{x}))^2]$) is completely determined by the overlap of $\mathbf{w}$ with itself and with the target direction $\mathbf{w}_\star$

$$Q = \frac{1}{D}\mathbf{w}\cdot\mathbf{w}\,,\; R = \frac{1}{D}\mathbf{w}\cdot\mathbf{w}_\star. \tag{1}$$

If the learning rate is scaled appropriately with the dimension D, the high-dimensional (large-D) limit of online stochastic gradient descent is given by a deterministic set of equations for Q and R:

$$\frac{d}{d\tau}\begin{bmatrix} Q(\tau) \\ R(\tau)\end{bmatrix} = \mathbf{F}[Q(\tau), R(\tau)], \tag{2}$$

where the continuous training "time" τ is the ratio of the number of samples seen to the dimension and $\mathbf{F}:\mathbb{R}^2 \to \mathbb{R}^2$ is a nonlinear function that depends on the learning rate, the loss function, and the link function $\sigma(\cdot)$ (Engel and van den Broeck, 2001; Ben Arous et al., 2022; Arnaboldi et al., 2023; Goldt et al., 2019; Saad and Solla, 1995). Integrating this update equation allows one to predict the evolution of the generalization error as more training data are provided to the algorithm. Despite the infinite dimensionality of the original optimization problem, only two dimensions are necessary to capture the dynamics of generalization error.

The analysis of online perceptron learning can be extended to two layer neural networks with a small number of hidden neurons N,

$$f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} a_i\,\phi\left(h_i(\mathbf{x})\right) \quad h_i(\mathbf{x}) = \frac{1}{\sqrt{D}}\mathbf{w}_i\cdot\mathbf{x}\,,\; i\in\{1,...,N\}. \tag{3}$$

$$y(\mathbf{x}) = \sigma\left(h_1^\star(\mathbf{x}),...,h_K^\star(\mathbf{x})\right) \quad h_k^\star(\mathbf{x}) = \frac{1}{\sqrt{D}}\mathbf{w}_k^\star\cdot\mathbf{x}\,,\; k\in\{1,...,K\}. \tag{4}$$

In this setting with isotropic random data, the relevant summary statistics are the readout weights $\mathbf{a} \in \mathbb{R}^N$, along with **overlap matrices** $\mathbf{Q}\in\mathbb{R}^{N\times N}$ and $\mathbf{R}\in\mathbb{R}^{N\times K}$ with entries

$$Q_{ij} = \frac{1}{D}\mathbf{w}_i\cdot\mathbf{w}_j\,,\; R_{ik} = \frac{1}{D}\mathbf{w}_i\cdot\mathbf{w}_k^\star \tag{5}$$

For this system, we can track the gradient descent dynamics for $\mathbf{a}$, $\mathbf{Q}$, and $\mathbf{R}$ through a generalization of Equation 2 (Goldt et al., 2019; Saad and Solla, 1995; Biehl and Schwarze, 1995; Goldt et al., 2020). This reduces the dimensionality of the dynamics from the $N + DN$ trainable parameters $\{a_i\}$, $\{w_j\}$ to $N + N^2 + NK$ summary statistics, which is significant when $D \gg N + K$. This reduction enables the application of analyses that cannot scale to high dimensions,

for instance control-theoretic methods to study optimal learning hyperparameters and curricula (Mori et al., 2025; Mignacco and Mori, 2025). Recent works have also begun to study approximations to these summary statistics when the network width N is also large, as further dimensionality reduction if possible when $\mathbf{Q}$ and $\mathbf{R}$ have stereotyped structures (Montanari and Urbani, 2025; Arnaboldi et al., 2023).

Under what conditions is this reduction possible? Fundamentally, the summary statistics $\mathbf{a}$, $\mathbf{Q}$, and $\mathbf{R}$ are sufficient to determine the network's performance so long as the preactivations h_i and $h_k^\star$ are approximately Gaussian. Thus, one can relax the assumption that the inputs $\mathbf{x}$ are exactly Gaussian so long as a central limit theorem applies to h_i and $h_k^\star$ (Goldt et al., 2019, 2020). Moreover, one can allow for correlations between the different input dimensions so long as h_i and $h_k^\star$ remain Gaussian. If $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \Sigma$, with a modification of the definition of the overlaps to $Q_{ij} = \frac{1}{D}\mathbf{w}_i \cdot \Sigma \mathbf{w}_j$ and $R_{ik} = \frac{1}{D}\mathbf{w}_i \cdot \Sigma \mathbf{w}_k^\star$ a similar reduction applies (Arnaboldi et al., 2023). One can even consider extensions to plasticity rules other than stochastic gradient descent. For example, online node perturbation leads to a different effective dynamics for the same set of summary statistics (Hara et al., 2011, 2013).

How could the overlaps $\mathbf{Q}$ and $\mathbf{R}$ be accessed from measurements of neural activity? And, in the absence of detailed knowledge of a teacher network, how could one identify the relevant overlaps? Under the simple structural assumptions of these models, one could estimate the overlaps from covariances of network activity across stimuli, i.e., with isotropic inputs one has $\mathbb{E}_{\mathbf{x}}[h_i h_k^\star] = R_{ik}$ and $\mathbb{E}_{\mathbf{x}}[h_i h_j] = Q_{ij}$. Moreover, one can in some cases detect this underlying low-dimensional structure by examining the principal components of the learning trajectory (Ben Arous et al., 2023). However, more theoretical work is required in this vein.

3.2 Learning in wide and deep neural networks

Another strategy to reduce the complexity of multilayer deep neural networks is to analyze the dynamics of learning in terms of representational similarity matrices (kernels) for each hidden layer of the network. Consider, for example, a deep fully-connected network with input $\mathbf{x} \in \mathbb{R}^D$,

$$
\begin{aligned}
f(\mathbf{x}, t) &= \frac{1}{\gamma\sqrt{N}} \sum_{i=1}^{N} w_i(t)\phi(h_i^{(L)}(\mathbf{x}, t)), \\
h_i^{(\ell+1)}(\mathbf{x}, t) &= \frac{1}{\sqrt{N}} \sum_{j=1}^{N} W_{ij}^{(\ell)}(t)\phi(h_j^{(\ell)}(\mathbf{x}, t)), \quad \ell \in \{1, \dots, L+1\}, \\
h_i^{(1)}(\mathbf{x}, t) &= \frac{1}{\sqrt{D}} \sum_{j=1}^{D} W_{ij}^{(0)}(t) x_j,
\end{aligned}
\tag{6}
$$

where t denotes training time. Instead of using online stochastic gradient descent to train the weights as we did in the preceding section, suppose we use gradient flow to minimize the average error on a fixed set of training examples. Moreover, instead of considering a regime where the hidden layer width N is small relative to the input dimension D, let us now consider very wide networks with $N \gg D$ (Figure 1a).

What are the relevant summary statistics in this case? Applying the chain rule to the dynamics of the network outputs, one finds the differential equation

$$
\frac{d}{dt} f(\mathbf{x}, t) = -\mathbb{E}_{\mathbf{x}'} \sum_{\ell} G^{(\ell+1)}(\mathbf{x}, \mathbf{x}', t, t) \Phi^{(\ell)}(\mathbf{x}, \mathbf{x}', t, t) \frac{\partial \mathcal{L}}{\partial f(\mathbf{x}', t)}, \tag{7}
$$

where $\mathcal{L}$ is the loss function and $\mathbb{E}_{\mathbf{x}'}$ denotes expectation over the training dataset (Jacot et al., 2018; Lee et al., 2019; Bordelon and Pehlevan, 2023b). Here,

$$
\Phi^{(\ell)}(\mathbf{x}, \mathbf{x}', t, t') = \frac{1}{N} \sum_{i=1}^{N} \phi(h_i^{(\ell)}(\mathbf{x}, t)) \phi(h_i^{(\ell)}(\mathbf{x}', t')) \tag{8}
$$

are **representational similarity matrices**, and

$$
G^{(\ell)}(\mathbf{x}, \mathbf{x}', t, t') = \frac{1}{N} \sum_{i=1}^{N} g_i^{(\ell)}(\mathbf{x}, t) \tag{9}
$$

$$
g_i^{(\ell)}(\mathbf{x}', t'), \quad g_i^{(\ell)}(\mathbf{x}, t) \equiv \gamma\sqrt{N} \frac{\partial f(\mathbf{x}, t)}{\partial h_i^{(\ell)}(\mathbf{x}, t)}, \tag{10}
$$

are **gradient similarity matrices**, which respectively compare the hidden states $\phi(h_i^{(\ell)}(\mathbf{x}, t))$ and the gradient signals $g_i^{(\ell)}(\mathbf{x}, t)$ at each hidden layer ℓ for each pair of data points $(\mathbf{x}, \mathbf{x}')$ and each pair of training times (t, t'). Thus, as $\Phi^{(\ell)}$ and $G^{(\ell)}$ determine the dynamics of f, these matrices are suitable summary statistics of learning if they are low-dimensional relative to the set of synaptic weights, and if we can write down a closed set of equations for their dynamics.

First, it is easy to see that the criterion of dimensionality reduction requires that the number of training examples P is much less than the network width N, as the number of similarity matrix elements and the number of synaptic weights are of order P^2 and N^2, respectively. Second, it turns out that one can close the equations for $\Phi^{(\ell)}$ and $G^{(\ell)}$ provided that the width is large and that the synaptic weights start from an uninformed initial condition (i.e., Gaussian random matrices) (Jacot et al., 2018; Lee et al., 2019; Yang and Hu, 2021; Bordelon and Pehlevan, 2023b). Depending on how weights and learning rates are scaled, one can obtain different types of large-width ($N \to \infty$) limits (Figure 1b). In the *lazy/kernel* limit where γ is constant, these representational similarity matrices are static over the course of learning (Jacot et al., 2018; Lee et al., 2019). However, an alternative scaling ($\gamma \propto \sqrt{N}$) can be adopted where these objects evolve in a task-dependent manner even as $N \to \infty$ (Figure 1c) (Yang and Hu, 2021; Bordelon and Pehlevan, 2023b).

While this provides a description of the training dynamics of a model under gradient flow, one can extend this description in terms of similarity matrices to other learning rules which use approximations of the backward pass variables $\tilde{g}_i^{(\ell)}(\mathbf{x}, t)$, which we called pseudo-gradients in Bordelon and Pehlevan (2023a). Such rules include Hebbian learning, feedback alignment, and direct feedback alignment (Hebb, 2005; Lillicrap et al., 2016; Nøkland, 2016). In this case, the relevant summary statistics to characterize the prediction dynamics of the network include the gradient-pseudogradient correlation, which measures the alignment between

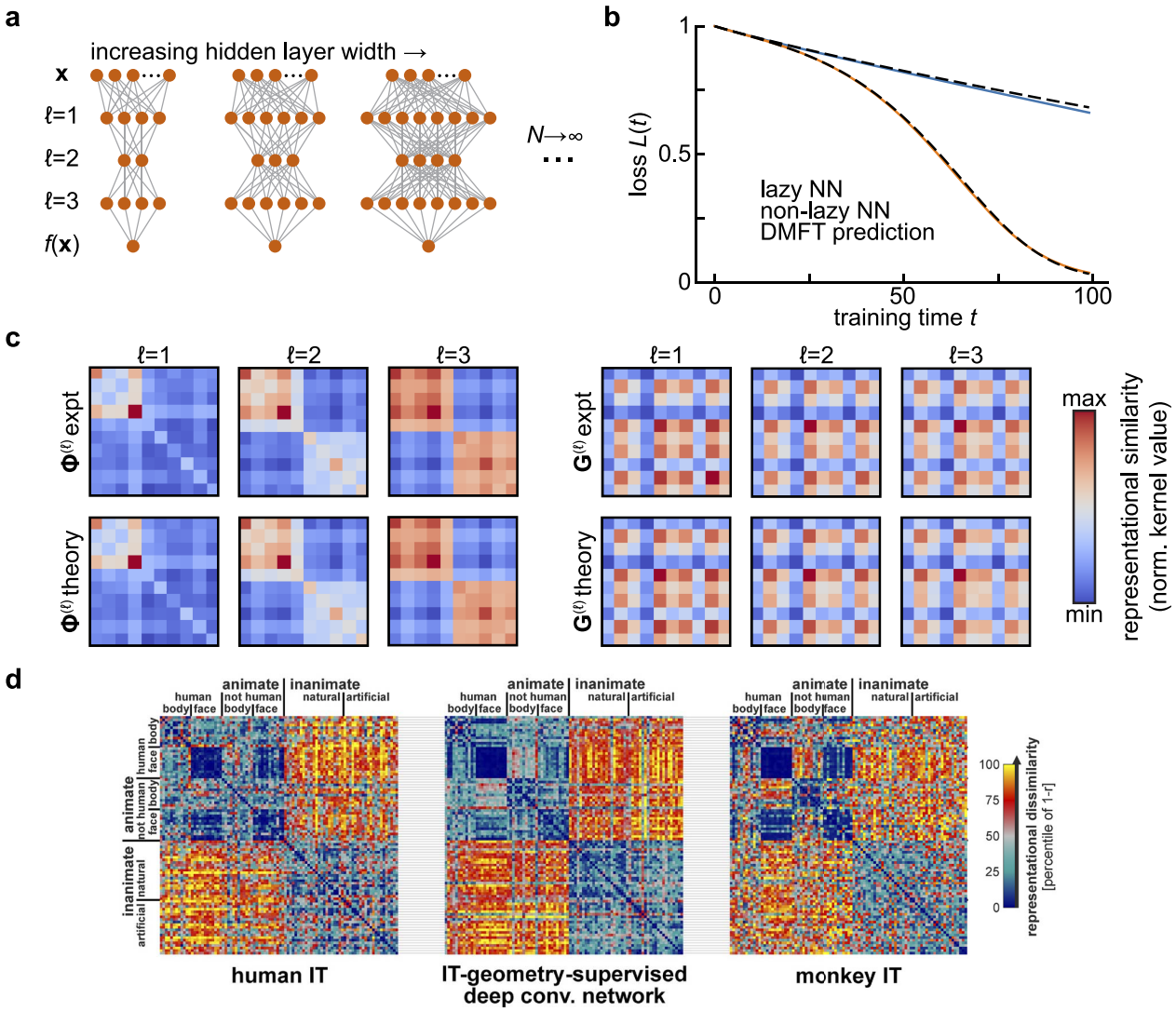

FIGURE 1
Representational similarity kernels in wide neural network models and in the brain. **(a)** Diagram of the infinite-width limit of a deep feedforward neural network. For a fixed input and output dimension, one considers a sequence of networks of increasing hidden layer widths. **(b)** Predicting the performance of width-2,500 fully-connected networks with three hidden layers and tanh activations over training using the dynamical mean-field theory described in Section 3. Networks are trained on a synthetic binary classification dataset of 10 examples, with 5 examples assigned each class at random. This leads to block structure in the final representations. Adapted from Bordelon and Pehlevan (2023b). **(c)** The summary statistics in the dynamical mean field theory for the network in **(b)** are representational similarity kernels [$\Phi^{(\ell)}$; *left*] and gradient similarity kernels (G^{ℓ}; *right*) for each layer. The top row shows kernels estimated from gradient descent training, and the bottom row the theoretical predictions. All kernels are shown at the end of training ($t = 100$). Adapted from Bordelon and Pehlevan (2023b). **(d)**. Comparing representational similarity kernels across models and brains. Here, similarity is measured using the Pearson correlation r, and the *dissimilarity* $1 - r$ is plotted as a heatmap. Kernels resulting from fMRI measurements of human inferior temporal (IT) cortex (*left*) and electrophysiological measurements of macaque monkey IT cortex (*right*) are compared with the kernel for features from a deep convolutional neural network after optimal re-weighting to match human IT (*center*). Adapted from Figure 10 of Khaligh-Razavi and Kriegeskorte (2014) with permission from N. Kriegeskorte under a CC-BY License.

the gradients used by the learning rule and the gradients that one would have used with gradient flow,

$$\tilde{G}^{(\ell)}(\mathbf{x}, \mathbf{x}', t, t') = \frac{1}{N} \sum_{i=1}^{N} g_i^{(\ell)}(\mathbf{x}, t) \tilde{g}_i^{(\ell)}(\mathbf{x}', t'), \tag{11}$$

as $\tilde{G}^{(\ell)}$ governs the evolution of the function output:

$$\frac{d}{dt} f(\mathbf{x}, t) = -\mathbb{E}_{\mathbf{x}'} \sum_{\ell} \tilde{G}^{(\ell+1)}(\mathbf{x}, \mathbf{x}', t, t) \Phi^{(\ell)}(\mathbf{x}, \mathbf{x}', t, t) \frac{\partial \mathcal{L}}{\partial f(\mathbf{x}', t)}. \tag{12}$$

4 Implications for neural measurements

The two example settings detailed in Section 3 show how the relevant summary statistics of learning depend on network architecture and learning rule. Theoretical studies are just beginning to map out the full space of possible summary statistics for different network architectures (Ben Arous et al., 2023; Goldt et al., 2019; Saad and Solla, 1995; Cui et al., 2023; Zavatone-Veth

and Pehlevan, 2021; Bordelon and Pehlevan, 2023b; Zavatone-Veth et al., 2022b; Saxe et al., 2013; Bordelon et al., 2025; Arnaboldi et al., 2023; van Meegen and Sompolinsky, 2025; Engel and van den Broeck, 2001; Zdeborová and Krzakala, 2016). Though details of the relevant summary statistics vary depending on the scaling regime and task—as illustrated by the examples above, where network width, training dataset size, and learning rule change the relevant statistics and their effective dynamics—they share broad structural principles. In all cases, summary statistics are defined by (weighted) averages over sub-populations of neurons within the network of interest, e.g., correlations of activity with task-relevant variables, or autocorrelations of activity within a particular layer in a deep network. Thanks to these common structural features, these varied theories of summary statistics have common implications for the analysis and interpretation of neuroscience experiments.

4.1 Benign sub-sampling

The summary statistics encountered in Section 3 are robust to sub-sampling thanks to their basic nature as averages over the population of neurons. These statistical theories in fact post a far stronger notion of benign sub-sampling: they result in neurons that are statistically exchangeable. This is highly advantageous from the perspective of long-term recordings of neural activity, as reliable measurement of summary statistics does not require one to track the exact same neurons over time. Instead, it suffices to measure a sufficiently large subpopulation on any given day. This obviates many of the challenges presented by tracking neurons over multiple recording sessions (Masset et al., 2022). Moreover, the variability and bias introduced by estimating summary statistics from a limited subset of relevant neurons can be characterized systematically (Kang et al., 2025; Bordelon and Pehlevan, 2024; Zavatone-Veth et al., 2022a). Taken together, these properties mean that summary statistics are relatively easy to estimate given limited neural measurements, provided that exchangability is not too strongly violated (Gao et al., 2017). We will return to this question in the Discussion, as a detailed analysis of the effects of non-identical neurons will be an important topic for future theoretical work. There are limits, however, to how far one can sub-sample. For instance, representational similarity kernels are more affected by small, coordinated changes in the tuning of many neurons than large changes in single-neuron tuning (Figure 2) (Kriegeskorte and Wei, 2021). Determining the minimum number of neurons one must record in order to predict generalization dynamics across learning will be an important subject for future theoretical work (Gao et al., 2017; Kriegeskorte and Wei, 2021).

4.2 Invariances and representational drift

Though by our definition the summary statistics mentioned in Section 3 are sufficient to predict the network's performance, they are not sufficient statistics for all properties of the neural code. In particular, in part because they arise from theories in which neurons become exchangable, they have many invariances. These invariances mean that individual tuning curves can change substantially without altering the population-level computation (Kriegeskorte and Wei, 2021). For instance, the representational similarity kernels are invariant under rotation of the neural code at each layer, enabling complete reorganization of the single-neuron code without any effect on behavior. Similarly, overlaps with task-relevant directions are invariant to changes in the null space of those low-dimensional projections. These invariances mean that focusing on summary statistics of learning sets a particular aperture on what aspects of representations one can assay.

At the same time, the invariances of summary statistics have important consequences for functional robustness. In particular, they are closely related to theories of representational drift, the seemingly puzzling phenomenon of continuing changes in neural representations of task-relevant variables despite stable behavioral performance (Rule et al., 2019; Masset et al., 2022). Many models of drift explicitly propose that representational changes are structured in such a way that certain summary statistics are preserved (Figure 2a) (Masset et al., 2022; Pashakhanloo and Koulakov, 2023; Qin et al., 2023). Identifying the invariances of the summary statistics sufficient to determine task performance can allow for a more systematic characterization of what forms of drift can be accommodated by a given network. Conversely, identifying the invariances of a representation once task performance stabilizes might suggest which summary statistics are relevant for the learning problem at hand.

4.3 Universality

An important lesson from the theory of high-dimensional statistics is that of *universality*: certain coarse-grained statistics are asymptotically insensitive to the details of the distribution. The most prominent example of statistical universality is the familiar central limit theorem: the distribution of the sample mean of independent random variables tends to a Gaussian as the number of samples becomes large. A broader class of universality principles arise in random matrix theory: the distribution of eigenvalues and eigenvectors of a random matrix often become insensitive to details of the distribution of the elements as the matrix becomes large. Most famously, the Marčenko-Pastur theorem specifies that the singular values of a matrix with independent elements have a distribution that depends only on the mean and variance of the elements (Marchenko and Pastur, 1967). In the context of learning problems, universality manifests through insensitivity of the model performance to details of the distributions of parameters or of features (Hu and Lu, 2022; Misiakiewicz and Saeed, 2024).

From the perspective of summary statistics, statistical universality can allow simple theories to make informative macroscopic predictions even if they do not capture detailed properties of single neurons. For instance, the mean-field description of the learning dynamics of wide neural networks introduced in Section 3 are universal in that they depend on the initial distribution of hidden layer weights only through its mean and variance, even though the details of that distribution will affect the distribution of weights throughout training (Figures 2b–d) (Golikov and Yang, 2022; Williams, 1996). Like the invariances to transformations of the neural population code mentioned

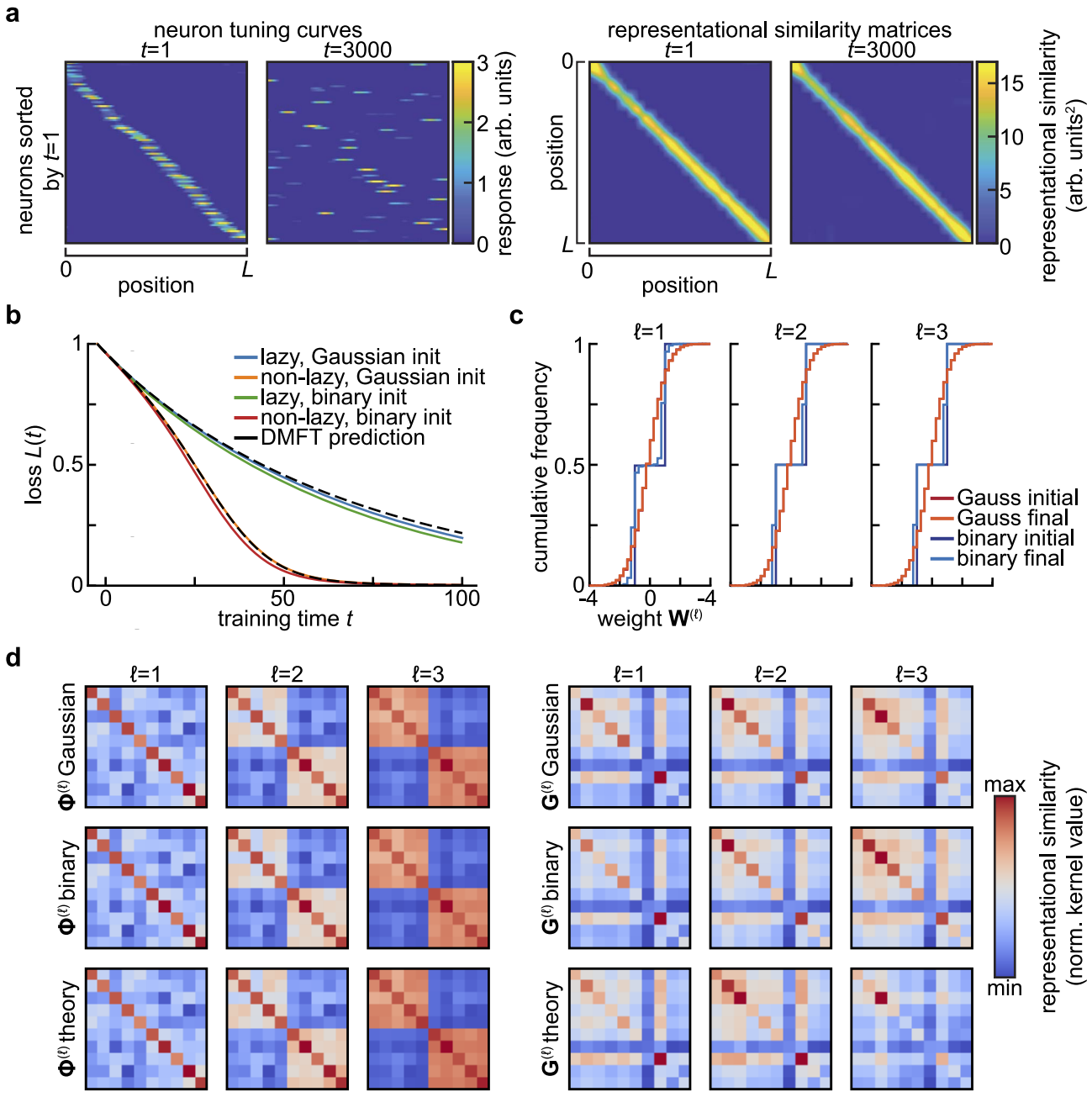

FIGURE 2
Invariance and universality in summary statistics. **(a)** Stable summary statistics despite drifting single-neuron responses. In Qin et al. (2023)'s model of representational drift, single neurons are strongly tuned to a spatial variable, yet their tuning changes dramatically over time (*left*). Despite this drift, the similarity of the population representations of different spatial positions remains nearly constant (*right*). Adapted from Figure 5e of Qin et al. (2023), of which C.P. is the corresponding author. **(b)** Universality of summary statistics in wide and deep networks with respect to the distribution of initial weights. Setting is as in Figures 1b, c, but also including a network for which the weights are initially drawn from $\{-1, +1\}$ with equal probability. Here, $N = 2,000$, and a different realization of the random task is sampled relative to Figures 1b, c, so the loss curves are not identical. **(c)** Cumulative distribution of weights at the start (*initial*) and end (*final*) of training for the networks shown in **(b)**. Note that the small change in the weight distributions for the Gaussian-initialized networks is not visible at this resolution, and that one expects the size of weight changes to scale with $1/\sqrt{N}$ (Bordelon and Pehlevan, 2023b). **(d)** Feature and gradient kernels at the end of training for the networks in **(b)**. No substantial differences are visible between networks initialized with different weight distributions.

before, this is nonetheless a double-edged sword: these universality properties mean that focusing on predicting performance commits one to coarse-graining away certain microscopic aspects of neural activity. Though these features are not required to predict macroscopic behavior, they may be important for understanding biological mechanisms.

5 Discussion

The core insight of the statistical mechanics of learning in neural networks is the existence of low-dimensional summary statistics sufficient to predict behavioral performance. We have reviewed how different summary statistics emerge depending on network architecture and task, how summary statistics might be estimated from experimental recordings, and what this perspective reveals about existing approaches to quantifying representational changes over learning. We now conclude by discussing complementary summary statistics of neural representations that arise from alternative desiderata, and future directions for theoretical inquiry.

A significant line of recent work in neuroscience aims to quantify neural representations and compare them across networks through analysis of representational similarity matrices $\Phi^{(\ell)}(\mathbf{x}, \mathbf{x}')$ (Kriegeskorte et al., 2008; Sucholutsky et al., 2024; Williams et al., 2021; Williams, 2024). Here, we see that these kernel matrices arise naturally as summary statistics of forward signal propagation in wide and deep neural networks (Figures 1c, d). At the same time, those results show that tracking only feature kernels is not in general sufficient to predict performance over the course of learning. One needs access also to coarse-grained information about the plasticity rule in the form of gradient kernels [either $G^{(\ell)}$ or $\tilde{G}^{(\ell)}$], and to information about the network outputs (for instance $\partial \mathcal{L}/\partial f$). More theoretical work is required to determine how to reliably estimate these gradient kernels from data, thereby providing a means to gain coarse-grained information about the underlying plasticity rule.

The summary statistics discussed here explicitly depend on the architecture and nature of plasticity in the neural network of interest, as they seek to predict its performance over learning. A distinct set of summary statistics arises if one aims to study what features of a representation are relevant for an *independently-trained* decoder. In this line of work, one regards the representation as fixed, rather than considering end-to-end training of the full network as we considered here. If the decoder is a simple linear regressor that predicts a continuous variable, the relevant summary statistics of the representation are just its mean and covariance across stimuli (Hu and Lu, 2022; Misiakiewicz and Saeed, 2024). Given a particular task, the covariance can be further distilled into the rate of decay of its eigenvalues and of the projections of the task direction into its eigenvectors (Hastie et al., 2022; Bordelon and Pehlevan, 2022; Canatar et al., 2021, 2024; Atanasov et al., 2024; Williams, 2024; Harvey et al., 2024; Bordelon et al., 2023). For categorically-structured stimuli, a substantial body of work has elucidated the summary statistics that emerge from assuming that one wants to divide the data according to a random dichotomy (Chung et al., 2018; Cohen et al., 2020; Bernardi et al., 2020; Farrell et al., 2022; Engel and van den Broeck, 2001; Zavatone-Veth and Pehlevan, 2022; Sorscher et al., 2022; Harvey et al., 2024).

The models reviewed here are composed of exchangeable neurons, which simplifies the relevant summary statistics and renders them particularly robust to sub-sampling. However, the brain has rich structure that can affect which summary statistics are sufficient to track learning and how those summary statistics may be measured. Biological neural networks are embedded in space, and their connectivity and selectivity is shaped by spatial structure (Khona et al., 2025; Chklovskii et al., 2002; Stiso and Bassett, 2018). Notably, many sensory areas are topographically organized: neurons with similar response properties are spatially proximal (Kandler et al., 2009; Murthy, 2011). Moreover, neurons can be classified into genetically-identifiable cell types (Zhang et al., 2023), which may play distinct functional roles during learning (Hirokawa et al., 2019; Fink et al., 2025). Future theoretical work must contend with these biological complexities in order to determine the relevant summary statistics of learning subject to these constraints.

Data availability statement

No experimental data were analyzed or generated in the preparation of this Perspective. Simulations of wide neural networks in Figures 1b, c, 2b–d following Bordelon and Pehlevan (2023b) are based on code available under an MIT License at https://github.com/Pehlevan-Group/dmft_wide_networks.

Author contributions

JZ-V: Conceptualization, Funding acquisition, Visualization, Writing – original draft, Writing – review & editing. BB: Conceptualization, Visualization, Writing – original draft, Writing – review & editing. CP: Conceptualization, Funding acquisition, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. JZ-V is supported by the Office of the Director of the National Institutes of Health under Award Number DP5OD037354. JZ-V is further supported by a Junior Fellowship from the Harvard Society of Fellows. BB is supported by a Google PhD Fellowship. CP is supported by NSF grant DMS-2134157, NSF CAREER Award IIS-2239780, DARPA grant DIAL-FP-038, a Sloan Research Fellowship, and The William F. Milton Fund from Harvard University. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence.

Acknowledgments

We are indebted to Nikolaus Kriegeskorte for sharing Figure 10 of Khaligh-Razavi and Kriegeskorte (2014), from which our Figure 1d is derived. We thank Paul Masset, Venkatesh Murthy, Farhad Pashakhanloo, and Ningjing Xia for helpful discussions and comments on previous versions of this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Author disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Arnaboldi, L., Stephan, L., Krzakala, F., and Loureiro, B. (2023). "From high-dimensional mean-field dynamics to dimensionless ODEs: a unifying approach to SGD in two-layers networks," in *Proceedings of Thirty Sixth Conference on Learning Theory, volume 195 of Proceedings of Machine Learning Research*, eds. G. Neu, and L. Rosasco (PMLR), 1199–1227.

Atanasov, A., Zavatone-Veth, J. A., and Pehlevan, C. (2024). Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*.

Ben Arous, G., Gheissari, R., Huang, J., and Jagannath, A. (2023). High-dimensional SGD aligns with emerging outlier eigenspaces. *arXiv preprint arXiv:2310.03010*.

Ben Arous, G., Gheissari, R., and Jagannath, A. (2022). "High-dimensional limit theorems for SGD: effective dynamics and critical scaling," in *Advances in Neural Information Processing Systems*, eds. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc.), 25349–25362.

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., and Salzman, C. D. (2020). The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* 183, 954–967.e21. doi: 10.1016/j.cell.2020.09.031

Biehl, M., and Schwarze, H. (1995). Learning by on-line gradient descent. *J. Phys. A Math. Gen.* 28:643. doi: 10.1088/0305-4470/28/3/018

Bordelon, B., Cotler, J., Pehlevan, C., and Zavatone-Veth, J. A. (2025). Dynamically learning to integrate in recurrent neural networks. *arXiv preprint arXiv:2503.18754*.

Bordelon, B., Masset, P., Kuo, H., and Pehlevan, C. (2023). "Loss dynamics of temporal difference reinforcement learning," in *Advances in Neural Information Processing Systems*, eds. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Curran Associates, Inc.), 14469–14496.

Bordelon, B., and Pehlevan, C. (2022). Population codes enable learning from few examples by shaping inductive bias. *Elife* 11:e78606. doi: 10.7554/eLife.78606

Bordelon, B., and Pehlevan, C. (2023a). "The influence of learning rule on representation dynamics in wide neural networks," in *The Eleventh International Conference on Learning Representations*.

Bordelon, B., and Pehlevan, C. (2023b). Self-consistent dynamical field theory of kernel evolution in wide neural networks. *J. Stat. Mech.* 2023:114009. doi: 10.1088/1742-5468/ad01b0

Bordelon, B., and Pehlevan, C. (2024). Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. *J. Statist. Mech.* 2024:104021. doi: 10.1088/1742-5468/ad642b

Canatar, A., Bordelon, B., and Pehlevan, C. (2021). Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nat. Commun.* 12:2914. doi: 10.1038/s41467-021-23103-1

Canatar, A., Feather, J., Wakhloo, A., and Chung, S. (2024). "A spectral theory of neural prediction and alignment," in *Advances in Neural Information Processing Systems*, 36.

Chklovskii, D. B., Schikorski, T., and Stevens, C. F. (2002). Wiring optimization in cortical circuits. *Neuron* 34, 341–347. doi: 10.1016/S0896-6273(02)00679-7

Chung, S., Lee, D. D., and Sompolinsky, H. (2018). Classification and geometry of general perceptual manifolds. *Phys. Rev. X* 8:031003. doi: 10.1103/PhysRevX.8.031003

Cohen, U., Chung, S., Lee, D. D., and Sompolinsky, H. (2020). Separability and geometry of object manifolds in deep neural networks. *Nat. Commun.* 11:746. doi: 10.1038/s41467-020-14578-5

Cui, H., Krzakala, F., and Zdeborova, L. (2023). "Bayes-optimal learning of deep random networks of extensive-width," in *Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research*, eds. A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (PMLR), 6468–6521.

Engel, A., and van den Broeck, C. (2001). *Statistical Mechanics of Learning*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139164542

Farrell, M., Bordelon, B., Trivedi, S., and Pehlevan, C. (2022). "Capacity of group-invariant linear readouts from equivariant representations: How many objects can be linearly classified under all possible views?," in *International Conference on Learning Representations*.

Fink, A. J., Muscinelli, S. P., Wang, S., Hogan, M. I., English, D. F., Axel, R., et al. (2025). Experience-dependent reorganization of inhibitory neuron synaptic connectivity. *bioRxiv.16.633450*. doi: 10.1101/2025.01.16.633450

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

Goldt, S., Advani, M., Saxe, A. M., Krzakala, F., and Zdeborová, L. (2019). "Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup," in *Advances in Neural Information Processing Systems*, 32. doi: 10.1088/1742-5468/abc61e

Goldt, S., Mézard, M., Krzakala, F., and Zdeborová, L. (2020). Modeling the influence of data structure on learning in neural networks: the hidden manifold model. *Phys. Rev. X* 10:041044. doi: 10.1103/PhysRevX.10.041044

Golikov, E., and Yang, G. (2022). "Non-gaussian tensor programs," in *Advances in Neural Information Processing Systems*, eds. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc.), 21521–21533.

Hara, K., Katahira, K., Okanoya, K., and Okada, M. (2011). Statistical mechanics of on-line node-perturbation learning. *IPSJ Online Trans.* 4, 23–32. doi: 10.2197/ipsjtrans.4.23

Hara, K., Katahira, K., Okanoya, K., and Okada, M. (2013). Statistical mechanics of node-perturbation learning for nonlinear perceptron. *J. Phys. Soc. Japan* 82:054001. doi: 10.7566/JPSJ.82.054001

Harvey, S. E., Lipshutz, D., and Williams, A. H. (2024). "What representational similarity measures imply about decodable information," in *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Stat.* 50, 949–986. doi: 10.1214/21-AOS2133

Hebb, D. O. (2005). *The Organization of Behavior: A Neuropsychological Theory*. London: Psychology press.

Hirokawa, J., Vaughan, A., Masset, P., Ott, T., and Kepecs, A. (2019). Frontal cortex neuron types categorically encode single decision variables. *Nature* 576, 446–451. doi: 10.1038/s41586-019-1816-9

Hu, H., and Lu, Y. M. (2022). Universality laws for high-dimensional learning with random features. *IEEE Trans. Inf. Theory* 69, 1932–1964. doi: 10.1109/TIT.2022.3217698

Jacot, A., Gabriel, F., and Hongler, C. (2018). "Neural tangent kernel: convergence and generalization in neural networks," in *Advances in Neural Information Processing Systems*, 31.

Kandler, K., Clause, A., and Noh, J. (2009). Tonotopic reorganization of developing auditory brainstem circuits. *Nat. Neurosci.* 12, 711–717. doi: 10.1038/nn.2332

Kang, H., Canatar, A., and Chung, S. (2025). Spectral analysis of representational similarity with limited neurons. *arXiv preprint arXiv:2502.19648*.

Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput. Biol.* 10, 1–29. doi: 10.1371/journal.pcbi.1003915

Khona, M., Chandra, S., and Fiete, I. (2025). Global modules robustly emerge from local interactions and smooth gradients. *Nature* 640, 155–164. doi: 10.1038/s41586-024-08541-3

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron* 93, 480–490. doi: 10.1016/j.neuron.2016.12.041

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:249. doi: 10.3389/neuro.06.004.2008

Kriegeskorte, N., and Wei, X.-X. (2021). Neural tuning and representational geometry. *Nat. Rev. Neurosci.* 22, 703–718. doi: 10.1038/s41583-021-00502-3

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., et al. (2019). "Wide neural networks of any depth evolve as linear models under gradient descent," in *Advances in Neural Information Processing Systems*, eds. H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc.). doi: 10.1088/1742-5468/abc62b

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nat. Commun.* 7:13276. doi: 10.1038/ncomms13276

Marchenko, V. A., and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik* 114, 507–536.

Masset, P., Qin, S., and Zavatone-Veth, J. A. (2022). Drifting neuronal representations: bug or feature? *Biol. Cyber.* 116, 253–266. doi: 10.1007/s00422-021-00916-3

Mézard, M., Parisi, G., and Virasoro, M. A. (1987). *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*. World Scientific Publishing Company. doi: 10.1142/0271

Mignacco, F., and Mori, F. (2025). A statistical physics framework for optimal learning. *arXiv preprint arXiv:2507.07907*.

Misiakiewicz, T., and Saeed, B. (2024). A non-asymptotic theory of kernel ridge regression: deterministic equivalents, test error, and GCV estimator. *arXiv preprint arXiv:2403.08938*.

Montanari, A., and Urbani, P. (2025). Dynamical decoupling of generalization and overfitting in large two-layer networks. *arXiv preprint arXiv:2502.21269*.

Mori, F., Mannelli, S. S., and Mignacco, F. (2025). "Optimal protocols for continual learning via statistical physics and control theory," in *The Thirteenth International Conference on Learning Representations*.

Murthy, V. N. (2011). Olfactory maps in the brain. *Annu. Rev. Neurosci.* 34, 233–258. doi: 10.1146/annurev-neuro-061010-113738

Nøkland, A. (2016). "Direct feedback alignment provides learning in deep neural networks," in *Advances in Neural Information Processing Systems*, 29.

Pashakhanloo, F., and Koulakov, A. (2023). "Stochastic gradient descent-induced drift of representation in a two-layer neural network," in *Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research*, eds. A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (PMLR), 27401–27419.

Qin, S., Farashahi, S., Lipshutz, D., Sengupta, A. M., Chklovskii, D. B., and Pehlevan, C. (2023). Coordinated drift of receptive fields in Hebbian/anti-Hebbian network models during noisy representation learning. *Nat. Neurosci.* 26, 339–349. doi: 10.1038/s41593-022-01225-z

Rule, M. E., O'Leary, T., and Harvey, C. D. (2019). Causes and consequences of representational drift. *Curr. Opin. Neurobiol.* 58, 141–147. doi: 10.1016/j.conb.2019.08.005

Saad, D., and Solla, S. A. (1995). On-line learning in soft committee machines. *Phys. Rev. E* 52:4225. doi: 10.1103/PhysRevE.52.4225

Saxe, A. M., McClelland, J. L., and Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

Sorscher, B., Ganguli, S., and Sompolinsky, H. (2022). Neural representational geometry underlies few-shot concept learning. *Proc. Nat. Acad. Sci.* 119:e2200800119. doi: 10.1073/pnas.2200800119

Steinmetz, N. A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., et al. (2021). Neuropixels 2.0: a miniaturized high-density probe for stable, long-term brain recordings. *Science* 372:eabf4588. doi: 10.1126/science.abf4588

Stiso, J., and Bassett, D. S. (2018). Spatial embedding imposes constraints on neuronal network architectures. *Trends Cogn. Sci.* 22, 1127–1142. doi: 10.1016/j.tics.2018.09.007

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., et al. (2024). Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*.

Sun, W., Winnubst, J., Natrajan, M., Lai, C., Kajikawa, K., Bast, A., et al. (2025). Learning produces an orthogonalized state machine in the hippocampus. *Nature* 640, 165–175. doi: 10.1038/s41586-024-08548-w

Vaidya, S. P., Li, G., Chitwood, R. A., Li, Y., and Magee, J. C. (2025). Formation of an expanding memory representation in the hippocampus. *Nat. Neurosci.* 28, 1510–1518. doi: 10.1038/s41593-025-01986-3

van Meegen, A., and Sompolinsky, H. (2025). Coding schemes in neural networks learning classification tasks. *Nat. Commun.* 16:3354. doi: 10.1038/s41467-025-58276-6

Watkin, T. L. H., Rau, A., and Biehl, M. (1993). The statistical mechanics of learning a rule. *Rev. Mod. Phys.* 65, 499–556. doi: 10.1103/RevModPhys.65.499

Williams, A. H. (2024). Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis. *bioRxiv. 2024-10*. doi: 10.1101/2024.10.23.619871

Williams, A. H., Kunz, E., Kornblith, S., and Linderman, S. (2021). "Generalized shape metrics on neural representations," in *Advances in Neural Information Processing Systems*, eds. M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc.), 4738–4750.

Williams, C. (1996). "Computing with infinite networks," in *Advances in Neural Information Processing Systems*, eds. M. Mozer, M. Jordan, and T. Petsche (MIT Press).

Yang, G., and Hu, E. J. (2021). "Tensor programs IV: feature learning in infinite-width neural networks," in *International Conference on Machine Learning* (PMLR), 11727–11737.

Zavatone-Veth, J. A., Canatar, A., Ruben, B. S., and Pehlevan, C. (2022a). Asymptotics of representation learning in finite Bayesian neural networks. *J. Statistical Mech.* 2022:114008. doi: 10.1088/1742-5468/ac98a6

Zavatone-Veth, J. A., and Pehlevan, C. (2021). "Depth induces scale-averaging in overparameterized linear Bayesian neural networks," in *Asilomar Conference on Signals, Systems, and Computers*, 55. doi: 10.1109/IEEECONF53345.2021.9723137

Zavatone-Veth, J. A., and Pehlevan, C. (2022). On neural network kernels and the storage capacity problem. *Neural Comput.* 34, 1136–1142. doi: 10.1162/neco_a_01494

Zavatone-Veth, J. A., Tong, W. L., and Pehlevan, C. (2022b). Contrasting random and learned features in deep Bayesian linear regression. *Phys. Rev. E* 105:064118. doi: 10.1103/PhysRevE.105.064118

Zdeborová, L., and Krzakala, F. (2016). Statistical physics of inference: thresholds and algorithms. *Adv. Phys.* 65, 453–552. doi: 10.1080/00018732.2016.1211393

Zhang, M., Pan, X., Jung, W., Halpern, A. R., Eichhorn, S. W., Lei, Z., et al. (2023). Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature* 624, 343–354. doi: 10.1038/s41586-023-06808-9

Zhong, L., Baptista, S., Gattoni, R., Arnold, J., Flickinger, D., Stringer, C., et al. (2025). Unsupervised pretraining in biological neural networks. *Nature* 2025, 1–10. doi: 10.1038/s41586-025-09180-y