ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

Unsupervised learning of mid-level visual representations

Giulio Matteucci[1], Eugenio Piasini[2] and Davide Zoccolan[2]

Abstract

Recently, a confluence between trends in neuroscience and machine learning has brought a renewed focus on unsupervised learning, where sensory processing systems learn to exploit the statistical structure of their inputs in the absence of explicit training targets or rewards. Sophisticated experimental approaches have enabled the investigation of the influence of sensory experience on neural self-organization and its synaptic bases. Meanwhile, novel algorithms for unsupervised and self-supervised learning have become increasingly popular both as inspiration for theories of the brain, particularly for the function of intermediate visual cortical areas, and as building blocks of real-world learning machines. Here we review some of these recent developments, placing them in historical context and highlighting some research lines that promise exciting breakthroughs in the near future.

Addresses

[1] Department of Basic Neurosciences, University of Geneva, Geneva, 1206, Switzerland

[2] International School for Advanced Studies (SISSA), Trieste, 34136, Italy

Corresponding author: Zoccolan, Davide (zoccolan@sissa.it)

(Matteucci G.)

(Zoccolan D.)

Current Opinion in Neurobiology 2024, **84**:102834

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 xxx

https://doi.org/10.1016/j.conb.2023.102834

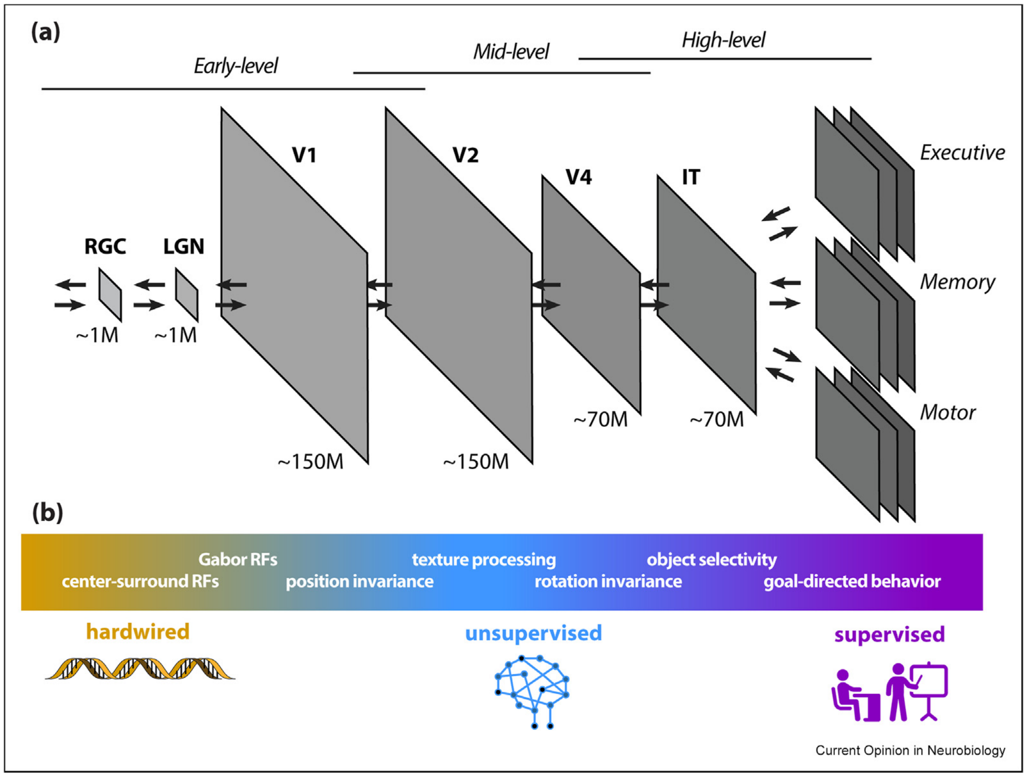
Introduction

Natural and artificial perceptual systems face the challenge of learning how to map the sensory environment into representations that are useful to guide perceptual decisions, motor behavior, and memory formation. Two primary methods exist for creating these maps. Learning can occur in the presence of a “teaching signal,” offering explicit feedback on the perceptual representations generated by the system. This feedback can range from precise information on the correct perceptual categorization of specific inputs (supervised learning) to a more generic rating of the overall perceptual processing outcome (reinforcement learning). Alternatively, learning can unfold in an unsupervised way, exploiting the inherent statistical structure of sensory experience without any explicit guidance (unsupervised learning) [1].

Reinforcement and supervised learning are thought to play a key role in the learning of sensorimotor transformations and the abstract cognitive representations guiding goal-directed behavior [2,3], as well as in fine-tuning sensitivity for perceptual tasks [4]. Instead, unsupervised learning has been suggested as the leading mechanism for the experience-dependent development of neuronal tuning from middle- (e.g. primary visual cortex - V1) to high-order areas (e.g. inferotemporal cortex - IT) along cortical sensory processing hierarchies [5]. This idea has two main motivations. First, unsupervised learning lends itself naturally to local implementations based on simple synaptic learning rules [6—15], obviating the need to transmit error feedback messages from high-order classification/decision centers to lower sensory areas. This problem, known as credit assignment, is solved by backpropagation in artificial neural networks (NNs), but, despite recent efforts, there is still no evidence for a mechanism that could have this role in the sensory cortex [16]. Second, during the early life of an animal, the amount of “labeled” sensory data is likely too little to provide the amount of training samples that are required for supervised learning [17].

Unsupervised learning could therefore support the continuous adaptation of cortical sensory representations to sensory input statistics, acting as a bridge between the largely hard-wired and evolutionary-determined processing circuits of low-level areas [18] (e.g. the retina) and the categorial/conceptual representations learned under supervision in higher-order memory/decision centers (Figure 1). In the following, after providing a historical perspective, we review the most recent evidence in support of the role played by unsupervised learning in the visual cortex.

Figure 1



Learning across the visual processing hierarchy. (a) Schematic of the primate visual processing hierarchy. Each area is plotted so that its size is proportional to its cortical surface area; the approximate total number of neurons (both hemispheres) is shown (M = million) following [5]. The shade of gray reflects the level of processing (light-to-dark = early-to-high). **(b)** Gradient of relevance of different wiring principles across the primate visual hierarchy, with associated keywords indicating some of the most important perceptual processes involved at each level.

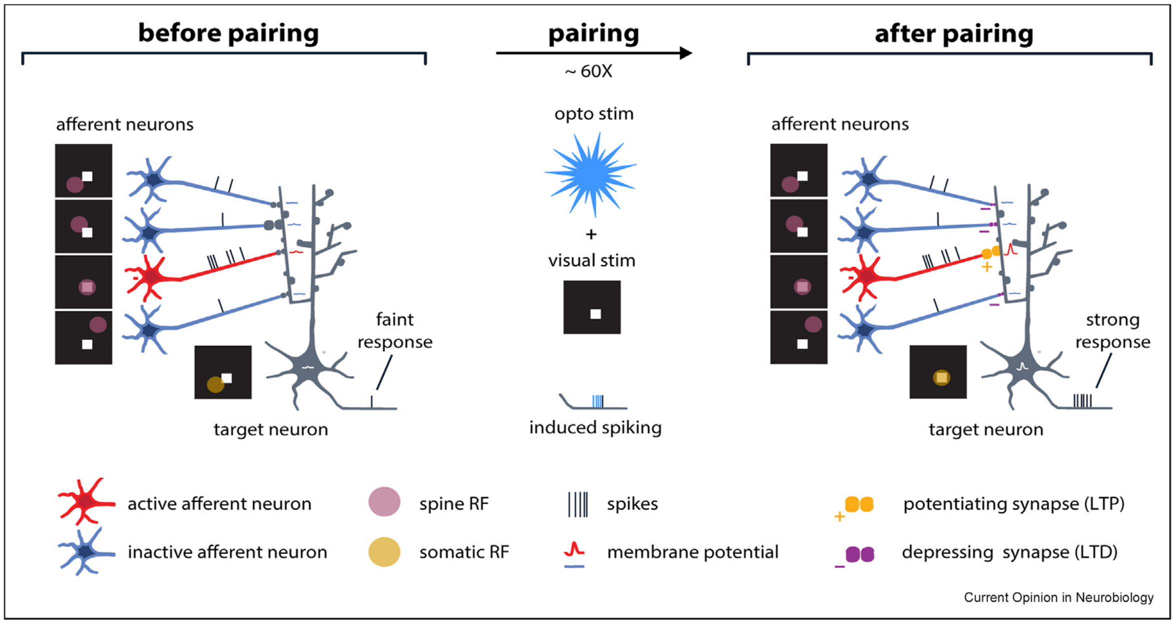
Synaptic mechanisms

The earliest and most influential hypothesis about how learning can take place in the brain can be traced back to Hebb's intuition [19] that, upon simultaneous activation, the strength of the synaptic connection between two neurons increases (Figure 3a). Hebb's speculation received neurobiological support from the discovery of long-term synaptic potentiation (LTP; initiated by a postsynaptic calcium influx via N-methyl-D-aspartate (NMDA) receptors) and long-term depression (LTD), followed by the realization that both LTP and LTD can depend upon the timing of presynaptic and postsynaptic spikes—a form of learning known as spike-timing-dependent plasticity (STDP) [20].

While many studies focused on connecting synaptic plasticity to memory, others focused on its role in experience-dependent sensory plasticity, e.g. demonstrating that ocular dominance plasticity in the visual cortex is prevented by blocking NMDA receptors [21]. Researchers successfully manipulated the response properties of single neurons by pairing visual stimulation with artificially triggered spiking activity, in agreement with Hebb's hypothesis [22]. In subsequent studies [23,24], by precisely controlling the time of postsynaptic spikes with respect to visual stimulation using in vivo patch clamp, it was possible to strengthen the responses of V1 neurons at specific visual field locations, thus obtaining a rapid unsupervised reshaping of neuronal receptive fields (RFs).

More recently, El-Boustani et al. [25] employed in vivo two-photon imaging and optogenetics to provide an unprecedented view of how synaptic plasticity underlies functional, unsupervised changes in sensory neurons (Figure 2). By consistently pairing visual stimuli with optogenetically-induced, temporally-precise spiking in V1 of awake mice, the authors triggered Hebbian plasticity, eventually shifting the receptive field of the targeted neuron towards the paired location. They also showed that spine volume changes correlated with the positional tuning of each synapse, indicating a transition from LTP (i.e. volume increase) to LTD (i.e. volume decrease) based on spine-RF center proximity to the paired location. LTP was found to be coordinated with gradual LTD of adjacent spines, consistent with a form of heterosynaptic plasticity acting in concert with Hebbian plasticity.

Figure 2



Hebbian plasticity can reshape visual RFs in vivo. chematic of the experiment by El Boustani et al. [25] showing a pyramidal neuron undergoing the visual-optogenetic stimulation pairing protocol. Left: initial condition. The neuron responds very weakly to visual stimulation of the target location indicated by the white rectangle. The efficacy of the synapse with the afferent neuron tuned to that location is low, as represented by its small volume. The somatic receptive field is offset with respect to the target location. After repeatedly pairing the presentation of the target stimulus with precisely timed optogenetic stimulation, which induces spiking in the target neuron, many spines undergo LTP or LTD. Right: outcome. The neuron responds strongly to visual stimulation of the target location. The efficacy of the synapse with the afferent neuron tuned to that location is increased while the others are reduced, as highlighted by corresponding changes in spine volume. The somatic receptive field now overlaps with the target location. LTP, long-term synaptic potentiation.

Another promising line of research focuses on understanding how top-down feedback, interacting with local circuit mechanisms, regulates synaptic plasticity, potentially helping the brain to solve the credit assignment problem [26]. For instance, acetylcholine can facilitate plasticity by disinhibiting NMDA-driven calcium entry in the dendritic segments of pyramidal neurons via activation of nicotinic receptors on specific inhibitory interneurons. This mechanism, likely critical in reinforcement learning, could also play a key role in unsupervised learning, where higher-level neurons might provide feedback to control the plasticity of relevant low-level features [26].

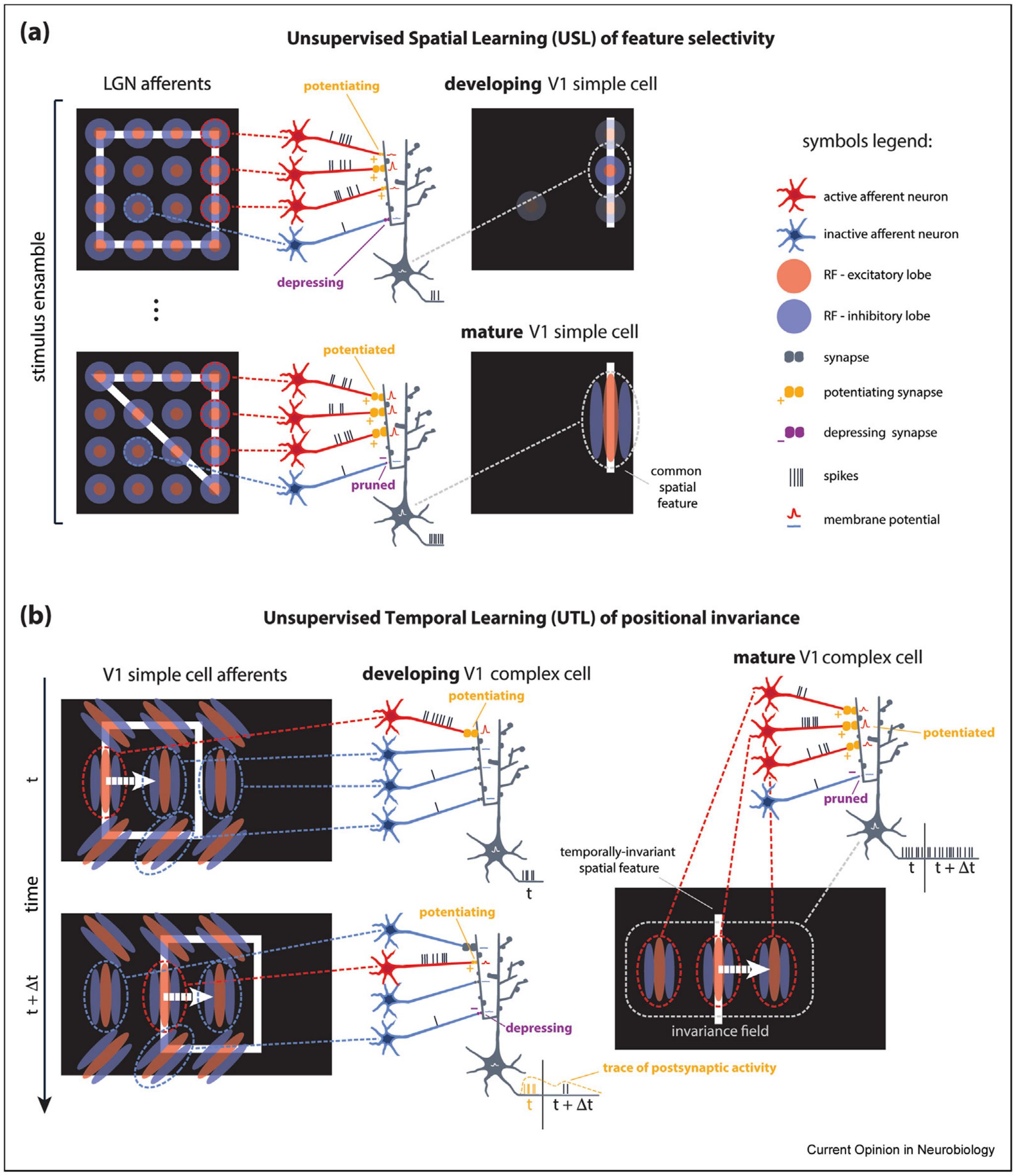
Theoretical foundation: From efficient coding to sparse coding and unsupervised temporal learning

From a theoretical perspective, it is useful to formulate high-level organizational principles that capture the emergent functional effect of learning in neural circuits. A notable example is the efficient coding principle, which, broadly speaking, postulates that neural processing is adapted to exploit the statistical structure of the natural environment. Efficient coding has been remarkably successful in generating compact theoretical descriptions of evolutionary-inherited principles of wiring in the sensory periphery (see references in Ref. [27]) as well as experience-dependent shaping of perceptual systems in central circuits [18].

Early insights into unsupervised learning of efficient codes came from the study of information maximization in artificial neural networks [11,12]. Later, Levy et al. showed that information and energetic efficiency can be incorporated by sparse codes, where only a few coding elements are active to represent each message [28]. Remarkably, Olshausen and Field showed that optimizing a neural code simultaneously for representational power and sparseness on natural images can lead to neural filters made of flanking, oriented subfields with alternating excitatory/inhibitory polarity, very similar to the RFs of V1 simple cells [29].

Sparseness is one of the key computational constraints that have been instantiated in the class of unsupervised learning approaches that we will refer to as *unsupervised spatial learning* (USL), because they rely only on the spatial statistics of natural input and ignore its temporal dimension. These approaches have also been applied to model V1 complex cells [30], the edge detector units

Figure 3



Learning selectivity and invariance through USL and UTL. **(a)** Schematic showing the hypothesized mechanism by which Hebbian USL could support the learning of V1 simple cells, starting from an LGN-like center-surround representation. Top: initial condition for the developing simple cell. One strong synapse drives the output of the neuron, as the somatic receptive field reflects that of the dominant afferent from lateral geniculate nucleus (LGN). Exposure to a stimulus ensemble containing a vertical bar as a common spatial feature induces repeated co-activation of several LGN-afferent neurons coding for different bar parts. Bottom: this exposure eventually causes Hebbian strengthening of the synapses of all co-activated afferents (and weakening of the others), leading to the formation of an oriented receptive field in the mature simple cell. **(b)** Schematic showing the hypothesized mechanism by which UTL, based on a trace rule, could support learning of position-invariant, oriented-edge detectors (i.e. complex cells), starting from a population of presynaptic simple cells. Left: the initial condition for the developing complex cell. One strong synapse drives the output of the neuron, and the somatic receptive field reflects that of this dominant simple cell afferent. As a stimulus containing its preferred vertical edge drifts through the RF of the dominant afferent, the postsynaptic neuron responds strongly (time = t, top). As the stimulus keeps drifting to an adjacent position (time = t + Δt, bottom), hitting the RF of another, weakly connected afferent, the lingering trace of recent activity in the postsynaptic neuron enables potentiation of the synapse with the currently active afferent. Right: this exposure eventually causes the strengthening of the synapses of all simple-cell afferents along the trajectory of the stimulus, endowing the neuron with a selectivity for the vertical bar at any point along the trajectory (i.e. position-invariant encoding). USL, unsupervised spatial learning; UTL, unsupervised temporal learning; RF, receptive field.

that encode orientation in a (locally) position-tolerant manner [31]. Transformation tolerance, however, has been more commonly modeled as the result of unsupervised temporal learning (UTL), namely those algorithms that exploit the full spatiotemporal structure of an incoming data stream. UTL exploits the temporal continuity of visual experience, where the natural tendency of different appearances (or views) of the same visual features to occur nearby in time is used to factor out feature identity (e.g. edge orientation) from other faster-varying, lower-level visual attributes (e.g. edge position).

The first instantiation of UTL can be traced back to Földiák's introduction of a "trace rule" as a variation of classical Hebbian learning, where the strengthening of a synapse is not merely proportional to the instantaneous firing of the presynaptic and postsynaptic units, but to the presynaptic activity and to a record of the recent history (a "memory trace") of postsynaptic activity [10]. This allows the postsynaptic cell to potentiate its synapses with presynaptic units that are activated sequentially over time (e.g., because a given visual feature sweeps through the visual field), thus inheriting their same selectivity while acquiring larger invariance (e.g. to translation or scaling; Figure 3b). Földiák's learning scheme successfully produced complex cell representations starting from the existence of a bank of simulated simple cells, as several later NN models based on UTL did [32–36]. Another family of UTL models produces complex cell RFs directly from the pixel (i.e. "retinal") representation, thus simultaneously learning shape selectivity and invariance. The most influential of such approaches is slow feature analysis (SFA), an unsupervised learning algorithm designed to extract slowly varying features from temporally varying input signals [37]. This approach finds a transformation of the input data that yields output features with minimal temporal variation, thus maximizing the "slowness" of the output signal. When applied to natural videos [38] or patterns of simulated spontaneous retinal activity [39], SFA produces response properties that are similar to those of V1 complex cells. When instantiated in a feedforward hierarchical structure, the slowness principle, often in combination with sparseness, is able to extract higher-order properties of a scene, including those encoded in deeper areas of the ventral visual stream [36,40] and the hippocampus [41]. Interestingly, recent efforts investigating plausible neurobiological implementations of USL and UTL have shown that STDP not only enables unsupervised extraction of complex visual features from natural images in a multilayer feedforward spiking neural network [42], but can also approximate independent component analysis [43] or slow feature analysis [15] under suitable conditions.

Maximization of sparseness and maximization of slowness can be seen as opposite pulling forces on neuronal representations—the first pushing neurons to become as selective as possible, the second leading them to respond as persistently as possible over time [44]. Thus, the two main classes of USL and UTL algorithms nicely recapitulate the well-known trade-off between shape selectivity and transformation tolerance that is inherent in object vision [5,45]. A central question is whether it is the interplay between USL and UTL that accounts for the build-up of selective yet invariant representations along visual processing hierarchies, such as the primate ventral stream [5]. Recent theoretical efforts in this direction include studies integrating sparse coding with manifold learning and slow feature analysis [46] or with predictive coding [47]. The next sections summarize the experimental evidence supporting this hypothesis.

Unsupervised spatial learning: Experimental evidence

Historically, the hypothesis that neuronal representations are adapted to the statistics of the visual world was causally tested by a number of studies rearing animals in altered visual environments, inspired by the concomitant discovery of a postnatal critical period of maximal cortical plasticity [48,49]. It was found that, in primary visual cortex, many tuning properties (e.g. for orientation and direction) are already present at eye opening in most species [48] (ferrets being the exception [50]). Visual experience, however, seems necessary to sharpen this "innate" tuning and maintain it. In cat V1, for instance, monocular deprivation has a strong impact on the development of ocular dominance, and restricting early visual experience to a single orientation biases orientation preference (see references in Ref. [51]). USL models based on sparse coding [51] well account for these phenomena, although we still lack experimental proof of one of their key predictions, i.e. that the lack of experience with images containing lines and contours prevents the development of V1-like RFs [29,51]. Despite early reports that rearing kittens in visual environments made of sparse dots leads to some reduction in the number of orientation-tuned cells in V1 [52], more recent experiments carried out on ferrets did not find any impairment in the development of orientation selectivity [53].

Going beyond V1, processing of visual textures has been used to study whether perception of complex visual patterns and higher-order cortical representations are also consistent with USL. From a perceptual standpoint, texture patterns that are more informative about visual scenes have been found to be more salient than less informative ones [54–56], consistent with the efficient coding principle applied to high-order visual statistics [56]. This result, initially established in humans for black-and-white visual textures, has been recently extended to grayscale images [57] and rodent vision

[58]. Neural representations of visual textures in intermediate ventral stream areas (V2, V4) have also been found to be adapted to the spatial statistics of natural images [59,60]. However, it remains unclear whether these representations are the result of adaptation to sensory statistics during development or arise from a USL process unfolding at the evolutionary time scale. In the future, this question could be addressed by controlled-rearing studies.

This approach was successfully employed to investigate the development of high-order facial representations in newborn monkeys. Arcaro et al. showed that macaques deprived of face exposure during early development did not develop face-selective patches in the IT cortex [61]. Previous work [62,63], on the other hand, suggested that monkeys reared without face experience still possessed an innate ability to detect faces. However, Sugita et al. also demonstrated that experience plays a crucial role in shaping the monkeys' preference for the specific category of commonly experienced faces (e.g. monkeys' vs. humans' faces) [63]. Both Arcaro's and Sugita's results suggest the experience-dependency of at least some aspects of face processing in primates. Consistent with recent computational work [64], this could result from an unsupervised learning process, although, given the social importance of faces for primates, reinforcement processes could also play an important role.

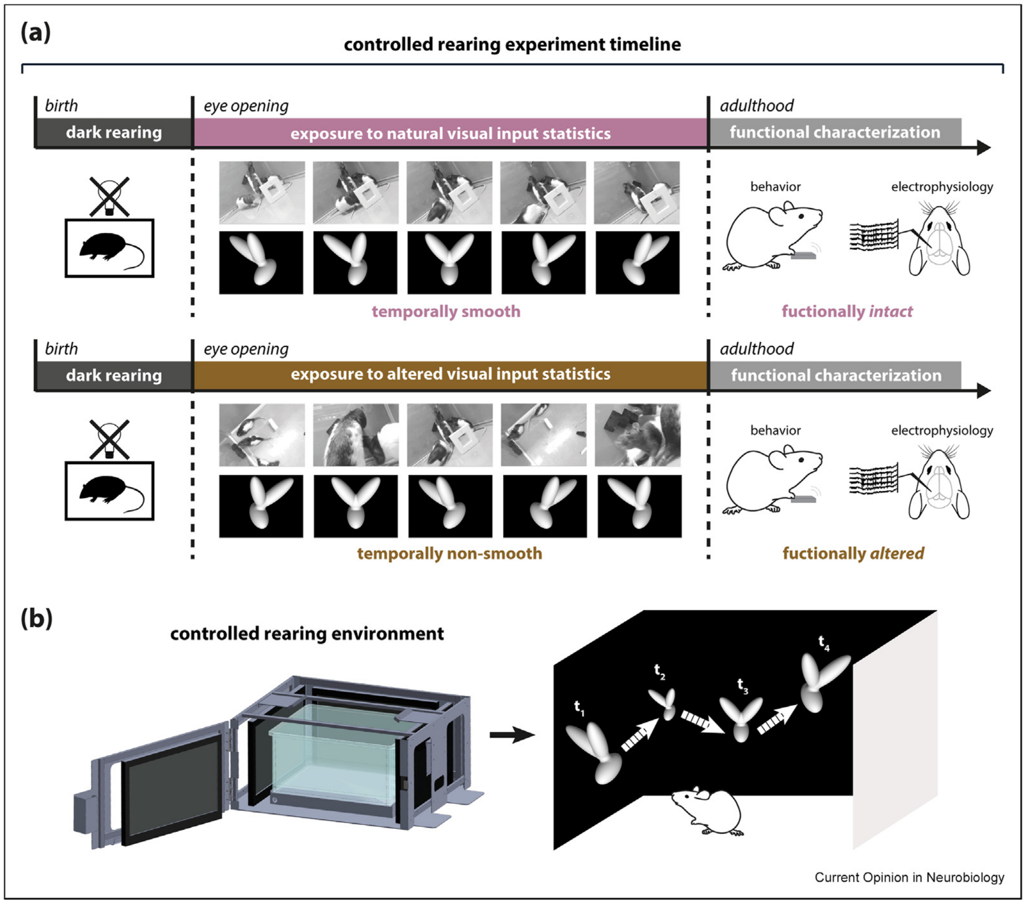
Unsupervised temporal learning: Experimental evidence

Theoretical UTL models prompted several psychophysical tests of the link between the spatiotemporal continuity of visual input and transformation tolerance in human vision. Wallis et al. [65,66] showed that it is possible to trick the visual system into considering two images of two different objects as being different views of the same object. This was achieved by passively exposing viewers to animations where one object, while undergoing viewpoint transformations (e.g. rotating in depth), also smoothly changed its identity (e.g. face A morphs into face B). A similar result was obtained for abrupt position changes produced by saccades when object identity was swapped in mid-saccade [67]. The ability to create such "false" invariances suggests that, in the adult brain, unsupervised learning mechanisms are continuously at work to associate temporally contiguous retinal images into the same perceptual representation. This conclusion is supported by other studies reporting better object recognition across in-depth rotations following a period of passive exposure to sequences of views of the object in close temporal proximity [68,69]. Notably, these studies found that object views do not need to be shown in sequential order (although this improves discrimination accuracy when few views per sequence are used [69]), but exposure to views presented in random order is enough to increase object recognition accuracy. This is in agreement with the findings of [67] but not entirely consistent with [65,66], where spatial consistency (i.e. smooth shape transformation) across temporally contiguous views was required to enable UTL for images of rotating face morphs. While the differences between these studies may be partly explained by the different types of stimuli used (faces vs. generic three-dimensional shapes), the importance of spatio-temporal consistency for UTL is further supported by recent studies on newborn chicks. This species develops more invariant object representations when reared with objects that rotate smoothly over time rather than changing views discontinuously [70] (Figure 4), with the amount of view invariance being negatively correlated with the speed of the rotation [71]. In summary, while temporal continuity seems to play an essential role in the learning of transformation tolerance, spatial continuity appears to have a facilitatory function (but see also [72] for a different view on the subject).

At the cortical level, the existence of UTL mechanisms shaping the tuning of visual neurons has been demonstrated by Li and DiCarlo [73–75], who exposed monkeys to altered associations between object identities across position and size changes. As a result of repeated exposures to such altered spatiotemporal statistics, neurons in IT weakened their original selectivity (e.g. preference for object A over B) at the swapped position/size and, in some cases, reversed it, thus either losing their tolerance across these transformations or even developing a "false" one. This reshaping of tolerance in IT takes place regardless of whether object transformation is abrupt (across saccades) or continuous (object A smoothly morphing into B while gradually changing size). The process develops over the course of a few hours of repeated exposures to the altered statistics and does not depend on the size of the reward or its temporal contingency with these exposures, nor on the actual engagement of the animal with the visual stimuli. This suggests that UTL is a fully unsupervised process (i.e. not requiring reward-based feedback) that allows the visual system to continuously adapt to the statistics of visual experience.

Recently, this UTL-based plasticity in monkey IT was shown to account for the reshaping of transformation tolerance reported in psychophysical studies. In Ref. [76], a Hebbian plasticity rule (designed to reduce the difference in neuronal responses to consecutive images) was implemented in a simulated population of IT neurons with realistic selectivity and tolerance. The model successfully captured the changes in tuning observed in IT following exposure to natural and altered spatiotemporal statistics [74]. Importantly, a simple linear readout of the simulated population accounted for the changes in object discrimination accuracy observed

Figure 4


(a)
controlled rearing experiment timeline
birth
dark rearing
eye opening
exposure to natural visual input statistics
adulthood
functional characterization
behavior
electrophysiology
temporally smooth
fuctionally intact
birth
dark rearing
eye opening
exposure to altered visual input statistics
adulthood
functional characterization
behavior
electrophysiology
temporally non-smooth
fuctionally altered
(b)
controlled rearing environment
t1
t2
t3
t4
Current Opinion in Neurobiology

Controlled rearing to test unsupervised learning theories. **(a)** Schematic of the timeline of a controlled rearing experiment. During the critical period of sensory development, animals are subjected to dark rearing from birth, except for repeated bouts of exposure to sensory stimuli with controlled statistics, e.g. in Ref. [77]: either natural movies or their frame-scrambled version; in Ref. [70]: either smoothly or nonsmoothly rotating objects. The neuronal and behavioral consequences of different “visual diets” can then be assessed and compared with theoretical predictions. **(b)** Example of a setup for controlled visual rearing. A transparent cage is surrounded by monitors, enabling immersive visual stimulation of newborn rodents. An example stimulus presentation is shown with a 3D object moving on the walls while rotating in depth and changing size (adapted from Ref. [77]).

in human subjects exposed to the same stimuli, thus providing a direct link between the psychophysical and neural implementation levels.

Another recent study [77] extended the investigation of UTL to a lower visual cortical area, focusing on tuning for much simpler features than visual objects and applying a fully ecological, unconstrained exposure to natural or altered spatiotemporal statistics (Figure 4). Newborn rats were reared, for the whole duration of the critical period, in controlled visual environments with either natural movies (control group) or their temporally scrambled version (experimental group). This controlled rearing led to a sizable reduction of the number of complex cells in V1 of the experimental rats and an impairment in their ability to encode orientation in a position-invariant way (compared to controls). This indicates that a form of UTL must be at work during postnatal development to leverage the spatiotemporal continuity of unconstrained visual experience and support the emergence of invariance in a visual area as low as V1. Crucially, such a learning process took place without the need for any explicit task or reward (which could have engaged general attentional/arousal mechanisms in Refs. [73–75]) or any ordered, sequential presentation of isolated stimuli (which could have helped the learning of relevant object representations in these earlier monkey studies). At the same time, the development of simple cells was not affected, and orientation tuning was equally sharp in both groups. This suggests that learning mechanisms based on USL, rather than UTL, underlie the development of shape selectivity, and that both forms of unsupervised learning are necessary to fully account for the development of visual cortical tuning [41,44,46].

Interestingly, a previous study in which adult rats were exposed to repeated swaps of grating orientation across spatial frequency (SF) changes found that this manipulation did not induce any change in orientation selectivity at the swapped SF [78]. In light of the findings reviewed above, demonstrating developmental UTL in rat visual cortex [77] and fast adult UTL in monkey IT cortex [73—75]), failure to observe UTL in Ref. [78] points to differences in the level of experience-dependent plasticity retained by sensory cortices across the lifespan. While high-order visual centers like IT retain their UTL-based plasticity well into adulthood (to constantly adapt high-order representations to the environment), the tuning of primary visual neurons for such basic features as oriented edges is likely more rigid during adulthood. These building blocks of visual representations are possibly learned once and for all during early postnatal development to remain stable after the closure of the critical period [48,49].

Rat visual cortex was also exploited in another recent study [79] to test whether visual representations become progressively slower across consecutive stages of the cortical object processing pathway. This should be the case if the role of UTL was to support the encoding of those features that, in the visual input, vary more slowly, like object identity [37]. However, such stimuli could engage adaptive and predictive processes, which could in principle fully counterbalance any increase in slowness of the cortical representation—for instance, by encoding preferentially "surprising" events, such as the appearance or disappearance of a stimulus, rather than its permanence within the field of view. In Ref. [79], rats were exposed to a set of natural movies, while neuronal responses were recorded from the anatomical progression of visual cortical areas (V1 → LM → LI → LL) that forms the rat homologue of the ventral stream [80]. Stimulus representations were indeed slower in higher-order areas as compared to V1, although differences were small ($\sim$50 ms). However, a much steeper hierarchy of temporal stability emerged when responses to movie segments containing isolated objects were considered, with representations in LI/LL being more than 1s slower than in V1. In addition, and consistent with previous findings in sensorimotor hierarchies in monkeys [81] and mice [82], the intrinsic timescale of the within-trial correlation of neuronal activity also increased along the cortical pathway, being $\sim$200 ms longer in LL than in V1. This may point to an increased role of recurrent and adaptive mechanisms along the hierarchy, which may support longer temporal receptive windows and may be beneficial for evidence accumulation and consistency of behavioral readouts [83]. Another intriguing possibility is that such extended lingering of postsynaptic activity may reflect the existence of mechanisms allowing higher-order neurons to deploy the "trace" learning rule at the base of UTL [10] over longer integration times. These findings were replicated by re-analyzing existing data recorded in mice [84], which additionally revealed a dependence of the coding timescales on the behavioral state of the animal.

Self-supervised learning on the rise

The last few years saw a surge in research papers connecting the machine-learning concept of "self-supervised learning" with neuroscience. Self-supervision is an increasingly popular form of unsupervised learning that creates meaningful data representations by solving a proxy task, with the intrinsic structure of input data acting as implicit supervision. Such proxy tasks might include reconstructing hidden image parts (masked autoencoders), restoring corrupted data (denoising autoencoders and diffusion models), generating new data samples (variational autoencoders and generative adversarial networks), or predicting the next data sample (contrastive predictive coding) [85]. Notably, some recent and influential work appeals directly to Barlow's redundancy reduction principle (i.e. discarding "repetitive" information, thereby efficiently encoding only the important aspects of the visual input) to arrive at nontrivial solutions for self-supervised learning problems [86].

Since the first wave of application of deep learning to explain brain representations, deep NNs trained with supervised learning methods have become the best models of high-level object vision in primates [87]. Recently, however, Zhuang et al. demonstrated that deep unsupervised contrastive embedding models can match or surpass supervised models in predicting neural activity in various ventral visual cortical areas [88]. Notably, these models mimic brain-like representations when trained solely on real-world child developmental data collected from head-mounted cameras. Furthermore, these techniques can be enhanced by sparse supervisory signals (such as occasional verbal labels) for semi-supervised learning using Local Label Propagation (LLP) [89]. LLP extrapolates labels for the entire dataset from a small number of available labels by inferring pseudo-labels for unlabeled images based on their proximity to labeled ones. Self-supervised models have also been claimed to outperform supervised methods in explaining visual representations in mice ventral and dorsal areas [90,91].

Other recent work reconnecting state-of-the-art unsupervised learning with neuroscience focused on explaining human brain activity. Konkle et al. demonstrated that unsupervised learning can extract category data from natural visual input and create representations as effective as category-supervised models in explaining human ventral stream functional magnetic resonance imaging (fMRI) representations [92]. This supports "domain-general" cognitive theories, which favor visual input statistics and universal learning mechanisms over innate biases in visual representation learning. In a similar vein, Choksi et al. demonstrated the potential of multimodal (i.e. visual-linguistic) self-supervised learning [93], as instantiated by the CLIP model (Contrastive Language-Image Pretraining; a model trained to associate images with textual descriptions [94]), in explaining the fMRI activity of the human hippocampus. CLIP has also been recently reported to contain units encoding highly abstract multimodal concepts [95], strikingly similar to medial temporal lobe neurons of human patients [96]. Other research effectively used unsupervised models to elucidate human psychophysical data, demonstrating that variational autoencoders could more accurately predict human perceptions and misperceptions of a complex perceptual feature such as gloss than their supervised counterparts [97].

Efferent copies of outgoing motor commands could also be instrumental in self-supervised learning. This idea goes back to classic developmental physiology studies [98], but has recently been explored with modern experimental and theoretical tools. Benucci, for instance, showed that integrating efferent copies of eye movements with visual inputs in convolutional neural networks enhances classification performance and provides robust visual representations invariant to saccadic shifts [99]. In a similar vein, Mineault et al. created a model of the primate dorsal visual pathway using a 3D convolutional NN to predict an agent's self-motion parameters from visual input [100]. Their network's responses strongly resembled those of primate and rodent visual neurons [100,101], underscoring the potential of exploring the interplay between self-initiated movements and the learning of sensory representations.

Recently, Halvagal et al. introduced the Latent Predictive Learning (LPL) plasticity rule, which merges a selectivity-building Hebbian learning term with an invariance-building predictive learning term [102]. LPL aims to learn stable latent object representations that predict future inputs, thereby forming unsupervised, disentangled, invariant representations in deep sensory networks. Furthermore, LPL recapitulates selectivity changes observed in primate inferotemporal cortex following altered visual experiences [73] as well as in rats reared in the absence of temporal continuity of the visual input [77].

Parallelly, Illing et al. formulated Contrastive Local and Predictive Plasticity (CLAPP), a self-supervised local learning rule for constructing deep hierarchical representations of images [103]. Drawing on spatial and temporal unsupervised learning principles and integrating predictive dendritic input, CLAPP constitutes a biologically plausible form of contrastive predictive learning using saccades as implicit time labels. CLAPP's demand for 'self-awareness' of saccades aligns it with the above-mentioned studies investigating self-motion-induced transformations for self-supervised learning. This rule aims at overcoming the need for back-propagation of error signals, providing a viable biological alternative for creating deep hierarchical representations that perform well in sensory classification tasks.

The recent studies reviewed in this section attest to the growing significance and practical application of various forms of self-supervised learning in providing normative accounts of sensory representations in the brain. Overall, we can discern an emerging trend: a shift from traditional supervised learning paradigms to a more biologically plausible framework of unsupervised learning, including self-supervised learning approaches.

Conclusions

Mid-level cortical representations are a good fit for unsupervised learning because they capture a large number of general-purpose, high-order features of a sensory data stream. By general-purpose, we mean that such representations are not specific to any task and are broadly relevant to perceptual experience. By high-order, we mean that they are not simple properties that can be easily computed directly from the elementary constituents of the data, like luminance or contrast from the amount of light impinging on photoreceptors in the retina. Therefore, learning such features without explicit supervision from the statistical structure of the world may be both possible, because of their generality, and necessary, because their large number and complexity make them intractable in a supervised setting (due to a lack of labeled data) or as a target for hardcoding by evolution (due to a genomic bottleneck [17]).

In this review, we have identified convergent trends across neuroscience and machine learning that leverage these intuitions to shed new light on the functioning of the sensory cortex at all levels—physiological, behavioral, and theoretical. With new experimental and computational techniques becoming available at a rapid pace, this approach promises significant breakthroughs in the coming years.

On the neurobiological front, emerging experimental techniques hold great promise for illuminating the plasticity rules at work in the brain. Genetic and molecular tools for studying plasticity in vivo have the potential to bridge synaptic changes to modifications of neuronal tuning and behavior [104]. For instance, super ecliptic pHluorin (SEP)-GluA1, a pH-sensitive green fluorescent protein (GFP)-tagged α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor subunit, allows direct imaging of AMPA receptor insertion produced by LTP during learning [105]. Similarly, genetic strategies using immediate early gene (IEG) promoters can tag active neurons, enabling selective expression of sensors and effectors in c-fos-positive neurons following specific experiences [106]. Such tools, already used to study the encoding of engrams [107, 108] and sensory stimuli [109], hold great promise for investigating the plasticity processes underlying unsupervised sensory learning. Finally, pairing different forms of imaging with spatially patterned and temporally precise optogenetic stimulation [110] could soon allow for investigating the unfolding of USL and UTL at the cellular and circuit level. For instance, it could be possible to simultaneously and repeatedly activate via optogenetic stimulation an ensemble of neurons tuned for different visual features [111], while tracking the neuronal and synaptic activity and looking for the emergence of a unit tuned for a composite, higher-order visual pattern made of such features. Alternatively, by repeatedly activating a set of V1 simple cells having similar orientation selectivity but RFs in slightly offset positions, it could become possible to directly observe the emergence of the invariant tuning of a complex cell. These experiments would finally allow to study directly the synaptic plasticity mechanisms at the base of the USL and UTL processes depicted in Figure 3, possibly down to the dendritic spine level.

On the computational front, we are witnessing the effectiveness of self-supervised learning in producing powerful artificial systems for the processing of sensory data. Increasingly, the neuroscience community will seize the opportunity to work with more biologically plausible systems that can replicate many sensory and cognitive computations of interest. In this context, machine learning and computational neuroscience researchers should join forces to bridge the gap between the abstraction level of global cost function minimization and the one of local learning rules. This is crucial because local learning rules are more directly interpretable in terms of synaptic mechanisms, allowing for practical testing in system neuroscience experiments. Starting from a common foundation, a renewed union of the study of biological plasticity and learning with modern machine learning will inspire more human-like and interpretable artificial intelligent systems and provide insights into the blueprints of the cortical circuits that compute the sensory representations underlying intelligent behavior.

Funding

This work was supported by the Italian Ministry of University and Research under the calls: PRIN 2022 (project no.2022XE8X9E to EP and project no. 2022WX3FM5 to DZ); Partenariato PNRR PE8 (project AGE-IT, Spoke 2 to DZ) and PRO3 (project NEMESI to DZ).

Credit author statement

Giulio Matteucci: Conceptualization, Visualization, Writing – original draft, Writing – review & editing. Eugenio Piasini: Conceptualization, Visualization, Writing – original draft, Writing – review & editing, Funding acquisition. Davide Zoccolan: Conceptualization, Visualization, Writing – original draft, Writing – review & editing, Funding acquisition.

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

No data was used for the research described in the article.

Acknowledgements

We would like to acknowledge Sami El-Boustani and Dimokratis Karamanlis for their insightful feedback on the original draft of the manuscript.

References

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

* of special interest
** of outstanding interest

1. Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, Clopath C, Costa RP, de Berker A, Ganguli S, *et al.*: **A deep learning framework for neuroscience**. *Nat Neurosci* 2019, **22**:1761–1770.

2. Botvinick M, Wang JX, Dabney W, Miller KJ, Kurth-Nelson Z: **Deep reinforcement learning and its neuroscientific implications**. *Neuron* 2020, **107**:603–616.

3. Yang GR, Molano-Mazón M: **Towards the next generation of recurrent network models for cognitive neuroscience**. *Curr Opin Neurobiol* 2021, **70**:182–192.

4. Dosher B, Lu Z-L: **Visual perceptual learning and models**. *Annu Rev Vis Sci* 2017, **3**:343–363.

5. DiCarlo JJ, Zoccolan D, Rust NC: **How does the brain solve visual object recognition?** *Neuron* 2012, **73**:415–434.

6. von der Malsburg Chr: **Self-organization of orientation sensitive cells in the striate cortex**. *Kybernetik* 1973, **14**:85–100.

7. Amari S: **Topographic organization of nerve fields**. *Bull Math Biol* 1980, **42**:339–364.

8. Kohonen T: **Self-organized formation of topologically correct feature maps**. *Biol Cybern* 1982, **43**:59–69.

9. Rumelhart DE, Zipser D: **Feature discovery by competitive learning**. *Cognit Sci* 1985, **9**:75–112.

10. Földiák P: **Learning invariance from transformation sequences**. *Neural Comput* 1991, **3**:194–200.

11. Linsker R: **Local synaptic learning rules suffice to maximize mutual information in a linear network**. *Neural Comput* 1992, **4**: 691–702.

12. Bell AJ, Sejnowski TJ: **An information-maximization approach to blind separation and blind deconvolution**. *Neural Comput* 1995, **7**:1129–1159.

13. Sirosh J, Miikkulainen R: **Topographic receptive fields and patterned lateral interaction in a self-organizing model of the primary visual cortex**. *Neural Comput* 1997, **9**:577–594.

14. de Sa VR, Ballard DH: **Category learning through multimodality sensing**. *Neural Comput* 1998, **10**:1097–1117.

15. Sprekeler H, Michaelis C, Wiskott L: **Slowness: an objective for spike-timing-dependent plasticity?** *PLoS Comput Biol* 2007, **3**: e112.

16. Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G: **Backpropagation and the brain**. *Nat Rev Neurosci* 2020, **21**: 335–346.

17. Zador AM: **A critique of pure learning and what artificial neural networks can learn from animal brains**. *Nat Commun* 2019, **10**:3770.

18. Daw N: *Visual development*. Springer Science+Business Media, LLC; 2013.

19. Hebb DO: *The organization of behavior: a neuropsychological theory*. John Wiley & Sons Inc.; 1949.

20. Magee JC, Grienberger C: **Synaptic plasticity forms and functions**. *Annu Rev Neurosci* 2020, **43**:95–117.

21. Kleinschmidt A, Bear MF, Singer W: **Blockade of "NMDA" receptors disrupts experience-dependent plasticity of kitten striate cortex**. *Science* 1987, **238**:355–358.

22. Frégnac Y, Shulz D, Thorpe S, Bienenstock E: **A cellular analogue of visual cortical plasticity**. *Nature* 1988, **333**: 367–370.

23. Meliza CD, Dan Y: **Receptive-field modification in rat visual cortex induced by paired visual stimulation and single-cell spiking**. *Neuron* 2006, **49**:183–189.

24. Pawlak V, Greenberg DS, Sprekeler H, Gerstner W, Kerr JN: **Changing the responses of cortical neurons from sub- to suprathreshold using single spikes in vivo**. *Elife* 2013, **2**, e00012.

25. ** El-Boustani S, Ip JPK, Breton-Provencher V, Knott GW, Okuno H, Bito H, Sur M: **Locally coordinated synaptic plasticity of visual cortex neurons in vivo**. *Science* 2018, **360**:1349–1354.

The authors demonstrate how Hebbian plasticity, induced by pairing visual stimuli with optogenetically-triggered spiking in vivo, can reshape a neuron's receptive field. The paper also shows how these functional changes are accompanied by predictable structural changes at the single spine level.

26. Roelfsema PR, Holtmaat A: **Control of synaptic plasticity in deep cortical networks**. *Nat Rev Neurosci* 2018, **19**:166–180.

27. Tesileanu T, Piasini E, Balasubramanian V: **Efficient processing of natural scenes in visual cortex**. *Front Cell Neurosci* 2022, **16**.

28. Levy WB, Baxter RA: **Energy efficient neural codes**. *Neural Comput* 1996, **8**:531–543.

29. Olshausen BA, Field DJ: **Emergence of simple-cell receptive field properties by learning a sparse code for natural images**. *Nature* 1996, **381**:607–609.

30. Karklin Y, Lewicki MS: **Emergence of complex cell properties by learning to generalize in natural scenes**. *Nature* 2009, **457**: 83–86.

31. Hubel DH, Wiesel TN: **Receptive fields of single neurones in the cat's striate cortex**. *J Physiol* 1959, **148**:574–591.

32. Wallis G: **Using spatio-temporal correlations to learn invariant object recognition**. *Neural Network* 1996, **9**:1513–1519.

33. Wallis G, Rolls ET: **Invariant face and object recognition in the visual system**. *Prog Neurobiol* 1997, **51**:167–194.

34. Einhäuser W, Kayser C, König P, Körding KP: **Learning the invariance properties of complex cells from their responses to natural stimuli**. *Eur J Neurosci* 2002, **15**:475–486.

35. Körding KP, Kayser C, Einhäuser W, König P: **How are complex cell properties adapted to the statistics of natural stimuli?** *J Neurophysiol* 2004, **91**:206–212.

36. Wyss R, König P, Verschure PFMJ: **A model of the ventral visual system based on temporal stability and local memory**. *PLoS Biol* 2006, **4**, e120.

37. Wiskott L, Sejnowski TJ: **Slow feature analysis: unsupervised learning of invariances**. *Neural Comput* 2002, **14**:715–770.

38. Berkes P, Wiskott L: **Slow feature analysis yields a rich repertoire of complex cell properties**. *J Vis* 2005, **5**:579–602.

39. Dähne S, Wilbert N, Wiskott L: **Slow feature analysis on retinal waves leads to V1 complex cells**. *PLoS Comput Biol* 2014, **10**, e1003564.

40. Franzius M, Wilbert N, Wiskott L: **Invariant object recognition and pose estimation with slow feature analysis**. *Neural Comput* 2011, **23**:2289–2323.

41. Franzius M, Sprekeler H, Wiskott L: **Slowness and sparseness lead to place, head-direction, and spatial-view cells**. *PLoS Comput Biol* 2007, **3**:e166.

42. Masquelier T, Thorpe SJ: **Unsupervised learning of visual features through spike timing dependent plasticity**. *PLoS Comput Biol* 2007, **3**:e31.

43. Savin C, Joshi P, Triesch J: **Independent component analysis in spiking neurons**. *PLoS Comput Biol* 2010, **6**, e1000757.

44. Lies J-P, Häfner RM, Bethge M: **Slowness and sparseness have diverging effects on complex cell learning**. *PLoS Comput Biol* 2014, **10**, e1003468.

45. Zoccolan D, Kouh M, Poggio T, Dicarlo J: **Trade-off between object selectivity and tolerance in monkey inferotemporal cortex**. *J Neurosci* 2007, **27**:12292–12307.

46. Chen Y, Paiton D, Olshausen B: **The sparse manifold transform**. In *Advances in neural information processing systems*. Curran Associates, Inc.; 2018.

47. Chalk M, Marre O, Tkačik G: **Toward a unified theory of efficient, predictive, and sparse coding**. *Proc Natl Acad Sci* 2018, **115**:186–191.

48. Espinosa JS, Stryker MP: **Development and plasticity of the primary visual cortex**. *Neuron* 2012, **75**:230–249.

49. Morishita H, Hensch TK: **Critical period revisited: impact on vision**. *Curr Opin Neurobiol* 2008, **18**:101–107.

50. White LE, Fitzpatrick D: **Vision and cortical map development**. *Neuron* 2007, **56**:327–338.

51. Hunt JJ, Dayan P, Goodhill GJ: **Sparse coding can predict primary visual cortex receptive field changes induced by abnormal visual input**. *PLoS Comput Biol* 2013, **9**, e1003005.

52. Blakemore C, Van Sluyters RC: **Innate and environmental factors in the development of the kitten's visual cortex**. *J Physiol* 1975, **248**:663–716.

53. Ohshiro T, Hussain S, Weliky M: **Development of cortical orientation selectivity in the absence of visual experience with contour**. *J Neurophysiol* 2011, **106**:1923–1932.

54. Tkačik G, Prentice JS, Victor JD, Balasubramanian V: **Local statistics in natural scenes predict the saliency of synthetic textures**. *Proc Natl Acad Sci* 2010, **107**:18149–18154.

55. Victor JD, Conte MM: **Local image statistics: maximum-entropy constructions and perceptual salience**. *JOSA A* 2012, **29**:1313–1345.

56. Hermundstad AM, Briguglio JJ, Conte MM, Victor JD, Balasubramanian V, Tkačik G: **Variance predicts salience in central sensory processing**. *Elife* 2014, **3**, e03722.

57. Tesileanu T, Conte MM, Briguglio JJ, Hermundstad AM, Victor JD, Balasubramanian V: **Efficient coding of natural scene statistics predicts discrimination thresholds for grayscale textures**. *Elife* 2020, **9**, e54347.

58. Caramellino R, Piasini E, Buccellato A, Carboncino A, Balasubramanian V, Zoccolan D: **Rat sensitivity to multipoint statistics is predicted by efficient coding of natural scenes**. *Elife* 2021, **10**, e72081.

59. Yu Y, Schmid AM, Victor JD: **Visual processing of informative multipoint correlations arises primarily in V2**. *Elife* 2015, **4**, e06604.

60. Ziemba CM, Freeman J, Movshon JA, Simoncelli EP: **Selectivity and tolerance for visual texture in macaque V2**. *Proc Natl Acad Sci* 2016, **113**:E3140–E3149.

61. ** Arcaro MJ, Schade PF, Vincent JL, Ponce CR, Livingstone MS: **Seeing faces is necessary for face-domain formation**. *Nat Neurosci* 2017, **20**:1404–1412.

This study reveals the necessity of early face exposure for the development of face-selective patches in the IT cortex of macaques, highlighting the role of experience in shaping high-order visual representations.

62. Sackett GP: **Monkeys reared in isolation with pictures as visual input: evidence for an innate releasing mechanism**. *Science* 1966, **154**:1468–1473.

63. Sugita Y: **Face perception in monkeys reared with no exposure to faces**. *Proc Natl Acad Sci* 2008, **105**:394–398.

64. Higgins I, Chang L, Langston V, Hassabis D, Summerfield C, Tsao D, Botvinick M: **Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons**. *Nat Commun* 2021, **12**:6456.

65. Wallis G, Backus BT, Langer M, Huebner G, Bülthoff H: **Learning illumination- and orientation-invariant representations of objects through temporal association**. *J Vis* 2009, **9**:6.

66. Wallis G, Bülthoff HH: **Effects of temporal association on recognition memory**. *Proc Natl Acad Sci* 2001, **98**:4800–4804.

67. Cox DD, Meier P, Oertelt N, DiCarlo JJ: **"Breaking" position-invariant object recognition**. *Nat Neurosci* 2005, **8**:1145–1147.

68. Liu T: **Learning sequence of views of three-dimensional objects: the effect of temporal coherence on object memory**. *Perception* 2007, **36**:1320–1333.

69. Tian M, Grill-Spector K: **Spatiotemporal information during unsupervised learning enhances viewpoint invariant object recognition**. *J Vis* 2015, **15**:7.

70. * Wood JN, Wood SMW: **The development of invariant object recognition requires visual experience with temporally smooth objects**. *Cognit Sci* 2018, **42**:1391–1406.

This controlled rearing study highlights that, as predicted by UTL, exposure to objects rotating smoothly over time enhances the development of invariant object recognition in newborn chicks.

71. Wood Justin N, Wood Samantha MW: **The development of newborn object recognition in fast and slow visual worlds**. *Proc R Soc B Biol Sci* 2016, **283**, 20160166.

72. Perry G, Rolls ET, Stringer SM: **Spatial vs temporal continuity in view invariant visual object recognition learning**. *Vis Res* 2006, **46**:3994–4006.

73. Li N, DiCarlo JJ: **Unsupervised natural experience rapidly alters invariant object representation in visual cortex**. *Science* 2008, **321**:1502–1507.

74. Li N, DiCarlo JJ: **Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex**. *Neuron* 2010, **67**:1062–1075.

75. Li N, DiCarlo JJ: **Neuronal learning of invariant object representation in the ventral visual stream is not dependent on reward**. *J Neurosci* 2012, **32**:6611–6620.

76. * Jia X, Hong H, DiCarlo JJ: **Unsupervised changes in core object recognition behavior are predicted by neural plasticity in inferior temporal cortex**. *Elife* 2021, **10**, e60830.

The authors show that human performance changes in object recognition induced by UTL are accounted for by a model of inferotemporal units, whose tuning is controlled by a Hebbian plasticity rule implementing UTL.

77. ** Matteucci G, Zoccolan D: **Unsupervised experience with temporal continuity of the visual environment is causally involved in the development of V1 complex cells**. *Sci Adv* 2020, **6**, eaba3742.

This study demonstrates that degrading the temporal continuity of visual experience during early postnatal life impairs the development of complex cells in the rat primary visual cortex while leaving simple cells unaffected. This underscores the role of UTL in the development of transformation tolerance.

78. Crijns E, Kaliukhovich DA, Vankelecom L, Op de Beeck H: **Unsupervised temporal contiguity experience does not break the invariance of orientation selectivity across spatial frequency**. *Front Syst Neurosci* 2019, **13**:22.

79. * Piasini E, Soltuzu L, Muratore P, Caramellino R, Vinken K, Op de Beeck H, Balasubramanian V, Zoccolan D: **Temporal stability of stimulus representation increases along rodent visual cortical hierarchies**. *Nat Commun* 2021, **12**:4448.

This study tests a long-standing prediction of unsupervised learning theories of visual cortical hierarchies–i.e. that deeper areas should encode slower features of a dynamic sensory stream. The authors show that this is the case as the time scales of both stimulus-driven responses and intrinsic neuronal activity increase across rat and mouse visual cortical areas.

80. Tafazoli S, Safaai H, De Franceschi G, Rosselli FB, Vanzella W, Riggi M, Buffolo F, Panzeri S, Zoccolan D: **Emergence of transformation-tolerant representations of visual objects in rat lateral extrastriate cortex**. *Elife* 2017, **6**, e22794.

81. Murray JD, Bernacchia A, Freedman DJ, Romo R, Wallis JD, Cai X, Padoa-Schioppa C, Pasternak T, Seo H, Lee D, *et al.*: **A hierarchy of intrinsic timescales across primate cortex**. *Nat Neurosci* 2014, **17**:1661–1663.

82. Runyan CA, Piasini E, Panzeri S, Harvey CD: **Distinct timescales of population coding across cortex**. *Nature* 2017, **548**:92–96.

83. Valente M, Pica G, Bondanelli G, Moroni M, Runyan CA, Morcos AS, Harvey CD, Panzeri S: **Correlations enhance the behavioral readout of neural population activity in association cortex**. *Nat Neurosci* 2021, **24**:975–986.

84. Siegle JH, Jia X, Durand S, Gale S, Bennett C, Graddis N, Heller G, Ramirez TK, Choi H, Luviano JA, *et al.*: **Survey of spiking in the mouse visual system reveals functional hierarchy**. *Nature* 2021, **592**:86–92.

85. Balestriero R, Ibrahim M, Sobal V, Morcos A, Shekhar S, Goldstein T, Bordes F, Bardes A, Mialon G, Tian Y, *et al.*: *A cookbook of self-supervised learning*. 2023.

86. Zbontar J, Jing L, Misra I, LeCun Y, Deny S: *Barlow twins: self-supervised learning via redundancy reduction*. 2021, https://doi.org/10.48550/arXiv.2103.03230.

87. Yamins DLK, DiCarlo JJ: **Using goal-driven deep learning models to understand sensory cortex**. *Nat Neurosci* 2016, **19**:356–365.

88. ** Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank MC, DiCarlo JJ, Yamins DLK: **Unsupervised neural network models of the ventral visual stream**. *Proc Natl Acad Sci* 2021, **118**.

This study shows that unsupervised contrastive embedding models rival supervised models in predicting primate ventral visual cortical activity, even when trained solely on real-world child developmental data.

89. Zhuang C, Ding X, Murli D, Yamins D: *Local label propagation for large-scale semi-supervised learning*. 2019, https://doi.org/10.48550/arXiv.1905.11581.

90. Bakhtiari S, Mineault P, Lillicrap T, Pack C, Richards B: **The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning**. In *Advances in neural information processing systems*. Curran Associates, Inc.; 2021:25164–25178.

91. Nayebi A, Kong NCL, Zhuang C, Gardner JL, Norcia AM, Yamins DLK: **Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation**. *PLoS Comput Biol* 2023, **19**, e1011506.

92. * Konkle T, Alvarez GA: **A self-supervised domain-general learning framework for human ventral stream representation**. *Nat Commun* 2022, **13**:491.

The authors show that a self-supervised model is able to account for human ventral stream fMRI representations, supporting theories favoring domain-general unsupervised mechanisms over innate category biases in visual learning.

93. Choksi B, Mozafari M, VanRullen R, Reddy L: **Multimodal neural networks better explain multivoxel patterns in the hippocampus**. *Neural Network* 2022, **154**:538–542.

94. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, *et al.*: *Learning transferable visual models from natural language supervision*. 2021, https://doi.org/10.48550/arXiv.2103.00020.

95. Goh G, NC, CV, Carter S, Petrov M, Schubert L, Radford A, Olah C: **Multimodal neurons in artificial neural networks**. *Distill* 2021, **6**:e30.

96. Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I: **Invariant visual representation by single neurons in the human brain**. *Nature* 2005, **435**:1102–1107.

97. Storrs KR, Anderson BL, Fleming RW: **Unsupervised learning predicts human perception and misperception of gloss**. *Nat Human Behav* 2021, **5**:1402–1417.

98. Held R, Hein A: **Movement-produced stimulation in the development of visually guided behavior**. *J Comp Physiol Psychol* 1963, **56**:872–876.

99. Benucci A: **Motor-related signals support localization invariance for stable visual perception**. *PLoS Comput Biol* 2022, **18**, e1009928.

100. * Mineault P, Bakhtiari S, Richards B, Pack C: **Your head is there to move you around: goal-driven models of the primate dorsal pathway**. In *Advances in neural information processing systems*. Curran Associates, Inc.; 2021:28757–28771.

The authors introduce a model of the primate dorsal visual pathway, which is trained, in a self-supervised way, to predict self-motion parameters from the visual input. The responses of units in the model closely align with those of neurons in the primate dorsal stream.

101. Matteucci G, Bellacosa Marotti R, Zattera B, Zoccolan D: **Truly pattern: nonlinear integration of motion signals is required to account for the responses of pattern cells in rat visual cortex**. *Sci Adv* 2023, **9**, eadh4690.

102. * * Halvagal MS, Zenke F: **The combination of Hebbian and predictive plasticity learns invariant object representations in deep sensory networks**. *Nat Neurosci* 2023, **26**: 1906–1915.

This study presents Latent Predictive Learning (LPL), a model that combines Hebbian and predictive plasticity to learn disentangled, invariant representations in deep sensory networks. LPL recapitulates changes seen in the primate and rodent cortex during altered visual experiences.

103. * Illing B, Ventura J, Bellec G, Gerstner W: *Local plasticity rules can learn deep representations using self-supervised contrastive predictions*. 2021, https://doi.org/10.48550/arXiv.2010.08262.

This study introduces Contrastive Local and Predictive Plasticity (CLAPP), a self-supervised local learning rule for deep hierarchical image representations that exploits the input temporal structure dictated by saccades as an implicit label.

104. Humeau Y, Choquet D: **The next generation of approaches to investigate the link between synaptic plasticity and learning**. *Nat Neurosci* 2019, **22**:1536–1543.

105. Roth RH, Cudmore RH, Tan HL, Hong I, Zhang Y, Huganir RL: **Cortical synaptic AMPA receptor plasticity during motor learning**. *Neuron* 2020, **105**:895–908.e5.

106. DeNardo L, Luo L: **Genetic strategies to access activated neurons**. *Curr Opin Neurobiol* 2017, **45**:121–129.

107. Choi J-H, Sim S-E, Kim J, Choi DI, Oh J, Ye S, Lee J, Kim T, Ko H-G, Lim C-S, *et al.*: **Interregional synaptic maps among engram cells underlie memory formation**. *Science* 2018, **360**: 430–435.

108. Abdou K, Shehata M, Choko K, Nishizono H, Matsuo M, Muramatsu S, Inokuchi K: **Synapse-specific representation of the identity of overlapping memory engrams**. *Science* 2018, **360**:1227–1231.

109. Tasaka G-I, Feigin L, Maor I, Groysman M, DeNardo LA, Schiavo JK, Froemke RC, Luo L, Mizrahi A: **The temporal association cortex plays a key role in auditory-driven maternal plasticity**. *Neuron* 2020, **107**:566–579.e7.

110. Panzeri S, Harvey CD, Piasini E, Latham PE, Fellin T: **Cracking the neural code for sensory perception by combining statistics, intervention, and behavior**. *Neuron* 2017, **93**:491–507.

111. Carrillo-Reid L, Yang W, Bando Y, Peterka DS, Yuste R: **Imprinting and recalling cortical ensembles**. *Science* 2016, **353**:691–694.