Critical periods support representation learning in a model of cortical processing

Ariane Delrocq, Zihan Wu, Guillaume Bellec, and Wulfram Gerstner

December 20, 2024

Abstract

Exposure of the brain to multiple stimuli drives the development of cortical representations, likely controlled by rules of synaptic plasticity. However, the type of developmental plasticity rules that lead to high-level representations of objects are unknown. Here we study a generalized Hebbian plasticity model that includes a predictive component. The learning rule uses only quantities that are locally available at the site of the synapse, is consistent with recent plasticity experiments in pyramidal neurons, and, as opposed to backpropagation learning, does not need a detailed feedback architecture. Our model shows that limiting plasticity in time to critical periods of development improves the quality and stability of sensory representation across different cortical areas described as layers of an artificial neural network. Our model achieves state-of-the-art performance for bio-plausible plasticity models on both an abstract hierarchical object database and a large image dataset designed for unsupervised learning.

1 Introduction

While visual input consists of high-dimensional retinal 'pixel' patterns, humans perceive the world as organized into objects like trees or coffee mugs. There are good reasons for the brain to transform sensory input patterns into more abstract representations. For example, different behavioral tasks might be learned more rapidly if the brain can rely on a common representation of the world. Along the ventral visual stream from early visual cortex V1 to higher visual areas in IT the level of abstraction increases from simple edge-detectors in V1 [1] to face-sensitive and object-sensitive neurons in IT [2, 3, 4]. Cortical representations are not genetically prewired but adapt during development [5, 6] and can be altered by changes in input statistics [7, 8, 9, 10, 11, 12, 13]. Importantly, these adaptation processes are most rapid during critical periods [14], i.e., specific time windows during which a given cortical area is highly plastic and particularly sensitive to specific aspects of sensory stimuli [15, 16]. Windows of critical plasticity have been found to be staggered across different areas, potentially to build up the elements necessary for a representation of relevant, but highly abstract features [17]. Language acquisition from auditory signals across multiple brain areas and time scales is a prominent example of this concept [18].

The opening and closing of critical periods involves multiple genetic, molecular, and cellular factors [17, 19] that enable the formation of new as well as changes of existing synapses. The rules of synaptic plasticity that enable abstract representations to emerge across several steps of the cortical processing stream remain difficult to grasp [4]. Classic models of Hebbian plasticity [20], where the change of a synapse depends only on the state of pre- (pre) and postsynaptic (post) neurons (and potentially on the momentary state of the synapse itself) have been useful to describe the development of receptive fields (RFs) in models of V1 [21, 22, 23, 24, 25, 26, 27, 28, 29], but are not sufficient to explain the increase of abstraction found across the ventral visual stream [4]. Currently, the most powerful models of object representation in the mammalian ventral stream are multilayer neural networks optimized by gradient descent with the backpropagation algorithm [30]-[36]. However, the backpropagation algorithm is biologically not plausible [37, 38], even though variants of the algorithm may increase some aspects of bio-plausibility [39]-[51]. A second path proposed to reconcile learning in deep networks with biology is the concept of predictive coding [52]-[58]. Despite the partial success of a recent study [59], it has overall been difficult to arrive, with Hebbian or predictive learning rules, at a hierarchical representation defined such that receptive fields in deeper cortical areas are both more complex *and* more useful than in the first areas. Indeed,

the performance of networks with biologically plausible Hebbian rules lacks far behind that of deep artificial neural networks trained with backpropagation [35].

However, recent plasticity experiments have reported observations that cannot be explained by standard Hebbian learning rules. In particular, synaptic plasticity at basal dendrites of pyramidal neurons depends on input arriving in the apical dendrite [60, 61], potentially conveying predictive signals from other neurons located in a different information processing pathway [60] or a higher cortical area [61]. Here we ask whether a generalized Hebbian learning rule that includes lateral predictive input into apical dendrites [62] can be used in combination with critical periods of plasticity to build abstract representations across a hierarchy of areas. From an abstract perspective, the lateral predictive input materializes the idea that within the *same* visual image obstructed parts should be predictable from their surround, consistent with experimental results [55, 63, 64] - but predictions should not be possible across separate images. In a broader sense, our approach is linked to contrastive predictive coding (CPC) in machine learning [65] -[72] but avoids the known problems of backpropagation [37, 38]. We show that critical periods of plasticity enable our generalized Hebbian learning rule with lateral input to achieve state-of-the-art performance on the classic image-base STL-10 for unsupervised learning [73] as well as on an abstract 'Random Hierarchy Model' designed to capture the sequential generation process of object-like stimuli [74].

2 Results

A generalized Hebbian learning rule with lateral inputs

We describe the visual stream from V1 to V2, V4, posterior, central, and inferior temporal cortex by a convolutional neural network with six layers where each network layer corresponds to one cortical area [33]. Neuron i in one of the areas has a firing rate

$$z_i = \rho(\sum_j w_{ij}\mathrm{pre}_j) \tag{1}$$

where pre_j is the rate of a neuron in the preceding area and w_{ij} is the feedforward weight that connects a given area with the following area in the processing stream. Compared to the cortex, the architecture of our model is simplified since we neglect the detailed microcircuits of excitatory and inhibitory neurons within each area. However, we take into account that, within each area, feedforward input arriving mainly at the basal dendrites of pyramidal neurons has a different role than feedback signals arriving mainly in the apical dendrites [75], either directly or indirectly via higher-order thalamus [38, 75, 76]. For the sake of simplification, we consider in the model that the major source of feedback stems from neurons in the same cortical area (Fig. 1), but we also considered model variants where feedback arrives from higher areas.

Inspired by classic results of synaptic plasticity, we describe the changes of a synapse from a presynaptic neuron j onto the basal dendrite of a postsynaptic neuron i by a Hebbian mechanism that depends on the firing rate pre_j of the presynaptic neuron as well as the state of the postsynaptic neuron [77, 78, 79]. In our model, the state of the postsynaptic neuron is characterized by two components. First, it depends on the local membrane potential post_i of the postsynaptic neuron which is a function $\mathrm{post}_i = f(a_i)$ of the total *feedforward* input $a_i = \sum_j w_{ij}\mathrm{pre}_j$ into the basal dendrite; second, it also depends on the total *lateral* input $\mathrm{lat}_i = \sum_k w_{ik}^{lat}\mathrm{z}_k$ into the apical dendrite caused by the activity z_k of other neurons in the same area

$$\Delta w_{ij} = \eta(t) \cdot \mathrm{lat}_i(t-\Delta) \cdot \mathrm{post}_i(t) \cdot \mathrm{pre}_j(t) \tag{2}$$

where Δ denotes a small delay of reception and integration of the apical dendritic input and $\eta(t)$ a learning rate. The lateral connections w_{ik}^{lat} change with a Hebbian learning rule as in [62].

Since we do not distinguish between excitatory and inhibitory neurons, all connections take positive or negative signs. Therefore the direction of change (potentiation $\Delta w_{ij} > 0$ or depression $\Delta w_{ij} < 0$, Eq. 2) of a model synapses w_{ij} onto the *basal* dendrite depends on the *apical* input of the postsynaptic neuron, consistent with experiments [61]. A learning rule consisting of multiplicative terms as in Eq. (2) can be derived from normative principles [62].

The synaptic plasticity rule formulated in Eq. (2) adjusts connectivity such that feedback input via the lateral connections matches the feedforward input. Thus, after learning, feedback input has a predictive role [55, 63, 64]. The learning rate $\eta(t)$ vanishes in our model, if the match between feedforward and feedback input is good, but is high if the match is not good (Methods). The

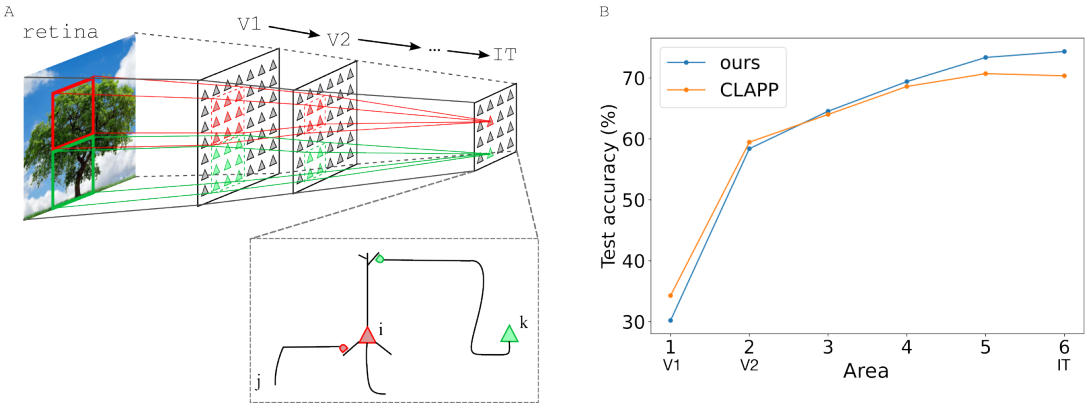

Figure 1: ***Network architecture and learning rule.*** *A. Layered artificial neural network representing areas V1, V2, V4, posterior, central, and inferior temporal (IT) cortex. Within each area, the network has lateral connections. If the image of a tree is projected onto the retina, a neuron (red triangle) with an RF in the red-framed square receives lateral input from other neurons (green triangles) with non-overlapping RFs (green square). Zoom: The plasticity of the feedforward connection* w_{ij} *from neuron j in the preceding area to neuron i (red) in IT depends on lateral input received by neuron i from other IT neurons k (green). B. Even though the learning rule works independently in each area, the classification performance for object recognition (linear readout) increases across areas, indicating increased usefulness of the internal representations along the model stream from V1 to IT. Our model (blue) is compared with a published model CLAPP [62].*

learning rate can therefore be considered as modulated by surprise [80, 81], i.e., learning is switched on if the expectation conveyed by lateral input does *not* match the actual input conveyed by the feedforward pathway. Such a switching process could be implemented by the coupling and decoupling of apical from basal dendrites at appropriate moments potentially controlled by thalamic input [82]. Appropriate normalization of synaptic weights is ensured in our model by switching the sign of plasticity during transitions between different images (Methods). We conjecture that transitions between different images would correspond to large saccadic eye movements which trigger brain signals [83, 84] that link views to each other[10, 85]. Alternative normalization schemes are also conceivable [59, 86]. To summarize, the generalized plasticity rule of Eq. (2) is, in a broad sense, in qualitative agreement with synaptic plasticity observed in cortical pyramidal neurons.

Compared to earlier work [51, 59, 61, 62] two aspects of our model are important. First, as explained further below, we implement critical periods [14, 15, 16, 17] and enable plasticity in the next area only after plasticity in the previous area has terminated. Second, in each cortical area lateral connections link localized patches of cortical neurons over relatively long distances in the retinotopic map. In particular, if an image such as a tree is presented to the retina, a neuron with a receptive field in the center of the image receives feedback input from other neurons with receptive fields located in other, non-overlapping parts of the image (Fig. 1A).

Experimental work has shown that the abstraction level of object representation increases across the visual stream from V1 to IT [4]. Consistent with earlier work in machine learning [62] we find in our model a similar increase in abstraction level measured by linear decodability of object classification (Fig. 1B), after training the network by presenting thousands of images of objects to the retina while running the learning rule of Eq. (2). We emphasize that during training we neither provide the identity of objects nor do we allow any backpropagation of information from IT to lower areas. Nevertheless, the usefulness of the representation, as measured by linear decodability, increases from model V1 to model IT.

Before we further investigate visual object recognition, we ask whether the learning rule with lateral prediction is able to extract abstract notions of 'objects' in an artificial data set with a hierarchical structure designed to mimic critical aspects of retinal images [74].

The plasticity rule represents abstract objects without BackPropagation

The recognition of images relies on two important aspects. First, most images contain redundant information. For example, we are able to identify a coffee mug from either its top or bottom segment alone (Fig. 2A). Second, the visual representation of an object in the retina can be interpreted as the result of a sequence of transformations from the raw object to the pixelized stimulation pattern

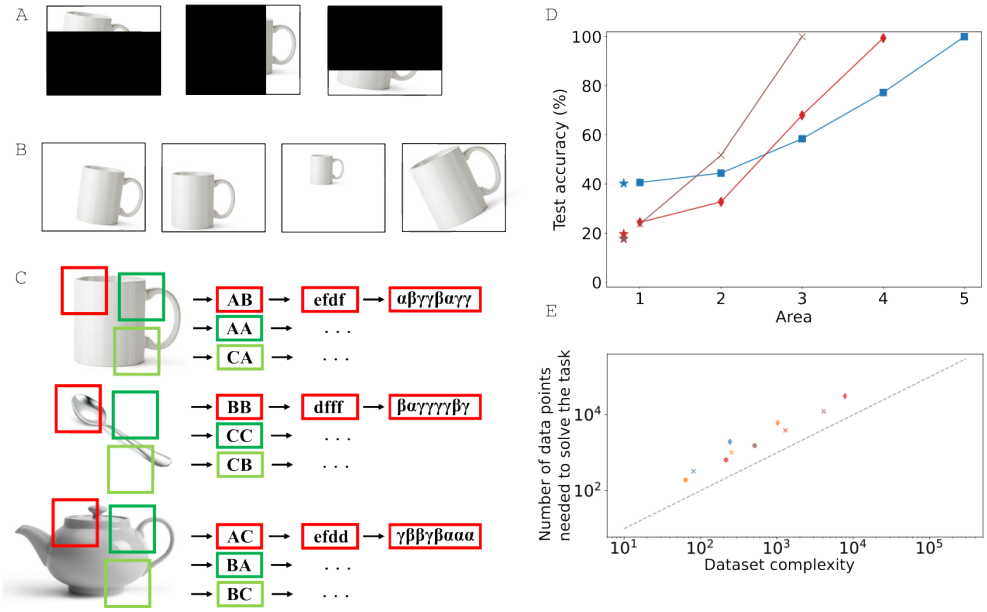

Figure 2: ***Learning exploits redundancy to extract the intrinsic structure in a high-dimensional dataset.*** *A. Most natural images contain redundant information. Small parts of an image are sufficient to roughly predict occluded parts of an image [63]. B. The same visual object may be projected onto the retina at different positions, inclinations, and sizes which can be interpreted as a multistep image rendering process [87]. C. The Random Hierarchy Model (RHM) generates abstract data by encoding a small number n_c of objects (e.g., mug, spoon, pot) in a multistep process of L steps [74]. At each step, an object (e.g., mug) is uniquely identified by several codewords (here $m = 3$ codewords per object are shown). At the next encoding step, each symbol A,B,C of the top level code is replaced by doublets of other symbols (d,e,f), which are then replaced by doublets of yet other symbols (α, β, γ), etc, yielding a final codeword with 2^L symbols. Two codewords of the same object together provide redundant information analogous to showing several image patches. D. Our model with L cortical areas, trained with pairs of symbol sequences of length 2^L applied to the retina is able to identify objects from the abstract codes generated by an RHM of L steps. A RHM coding depth of $L = 5$ (with $m = 3$ equivalent codes per symbol, blue) or $L = 4$, ($m = 6$, red) or $L = 3$ ($m = 8$, brown) is decoded in the cortical processing stream with 5, 4 or 3 areas, respectively. Test accuracies are measured by linear classification. Asterisks - linear classification applied to the raw code of length 2^L. E. Our model of bioplausible learning solves the dataset as efficiently as supervised backpropagation algorithms [74]: the number of samples needed for near-perfect representations of the dataset scales with the dataset complexity, defined as $n_c m^L$ (where m is the number the different doublets representing any class/symbol) even if this is only a small fraction of the total number $P_{max} = m^{(2^L)}$ of data points. Showing the minimum number of data points needed to solve different datasets: circles $L = 2$, crosses $L = 3$, diamonds $L = 4$; blue $m = 3$, orange $m = 4$, red $m = 6$, brown $m = 8$.*

('image rendering' [87]). For example, an object such as a coffee mug may appear on different scales, for a given scale in different positions, and for a given position in different orientations or viewing angles (Fig. 2B).

To mimic the sequential generation process of images, but without the inherent redundancy of images, the Random Hierarchy Model (RHM) [74] starts with n_c different objects where each object is represented at the most abstract level by m different codewords. Each codeword is a string of two symbols which are randomly selected from a longer list of symbols. In the example of Fig. 2C, the mug is represented by $m = 3$ different codewords of two symbols selected from a list of $v = 3$ possible symbols A, B, or C. Importantly, the assignment is random so that, e.g., the mug could be represented by any one of the codewords AA, AB, or CA and the teapot by any one of the codewords AC, BA, or BC. Therefore a symbol A in the first position of the string does not identify the object. At the next encoding step, each symbol is again encoded by m random combinations of two out of a second set of v symbols (say d,e,f) which in turn are again encoded by a further set of symbols (say α, β, γ), and so on. With combinations of two new symbols (selected from a potentially longer list of symbols) to replace a symbol from the previous stage, the total 'pixel-level' representation of an object after L sequential encoding steps is a string of length 2^L. Importantly, no small segment of the string of symbols can uniquely identify the original object - analogously to the difficulty of uniquely detecting an object from a small set of its pixels.

The Random Hierarchy Model was specifically designed to test backpropagation learning in deep networks [74]. Here we ask whether our general Hebbian plasticity rule with lateral predictions is able to extract useful representations of objects encoded with an L-step RHM in the absence of any backpropagation signals. To account for redundant information in images (Fig. 2C), we slightly modify the RHM and assume that each 'image' of an object consists of two of the allowed codewords of the object in the RHM.

In order to test the quality of the representation extracted by our model we try to predict the object label with a linear decoder from different areas of the model [33, 36]. We find that our network with the lateral predictive Hebbian rule can perfectly represent objects encoded by codewords of the RHM provided that the number of areas in the cortical processing stream is at least as large as the number L of encoding steps of the RHM (Fig. 2D). More generally, our learning rule leads to a near-perfect representation in area L of the processing stream if it is trained with a number of data samples that is slightly above the data complexity defined as $n_c \cdot m^L$ (Fig. 2E) where n_c is the number of objects and m^L is the number of different codewords for each object, – in agreement with the results found with full backpropagation in deep neural networks [74]. Thus, our network without any feedback connections and a generalized Hebbian learning rule acting locally at the synapse is able to perform as well on this data set as a model with full backpropagation. These findings suggest that our learning rule with lateral predictions is not specific and limited to visual processing, but generically useful for detecting low-dimensional 'object-like' structures in high-dimensional data.

Developmental learning is more efficient than traditional learning

In sensory pathways such as vision [15] or audition [16] synaptic plasticity was found to be largest during critical periods. It has been suggested that these critical periods are staggered sequentially in order to support the extraction of abstract features [17].

In order to test the hypothesis that sequential staggering of critical periods across the cortical hierarchy supports learning, we compare a sequential protocol of developmental learning with a control model where all areas learn in parallel (Fig. 3). We implement a developmental sequence of critical periods in our network model by training first V1 with our lateral-input dependent plasticity rule, while keeping the weights in other areas fixed at small, but random initial values. After a predefined number of training steps, the critical window for area V1 closes and weights to and within area V1 are kept henceforth at a fixed value. With the closure of the critical window in the first area, the critical period for the second area opens. This sequence ends with the critical period for area IT (Fig. 3B). For training we use the image dataset STL-10, designed for unsupervised learning [73].

We find that networks trained with sequential critical periods achieve representations that are better than a control network trained in the standard fashion, i.e. where all weights across all areas are learned simultaneously. More importantly, our experiments show that learning with a staggered sequence of critical periods is far more efficient in terms of resources, measured as the total number of synaptic changes across the network, than simultaneous learning (Fig. 3C). Note that in natural environments, visual input is abundant making not the amount of data, but rather potentially costly changes of synaptic connections the limiting factor.

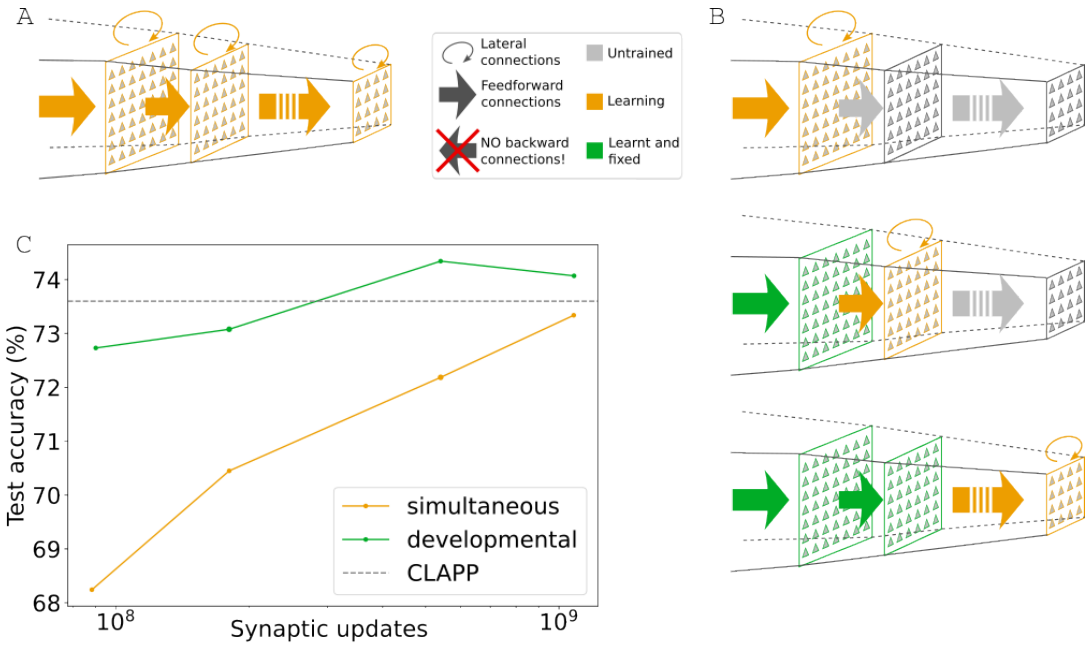

Figure 3: ***Critical periods improve the efficiency of learning.*** *A. Network architecture. Learning of feedforward connections depends on lateral input, but there are no backward connections. In a simultaneous learning paradigm, all connections change in parallel. B. Development learning is implemented as a sequence of critical periods (top to bottom) starting in V1 and finishing in IT. The orange color indicates connections to and inside the area that are amenable to learning during the critical period. C. Test accuracy for linear classification on STL10 as a function of synaptic updates. Developmental learning (green) reaches better performances with fewer synaptic updates than simultaneous learning (orange). For comparison, the performance of CLAPP as reported in [62] is indicated by the horizontal dashed line.*

3 Discussion

Critical periods of increased plasticity have been found across brain areas and animal species [14, 15, 16, 17, 18] which suggests that critical periods might be generically useful to adapt the neuronal representations of the outside world to the statistics of the environment. We found in a model of a cortical processing chain across multiple areas that learning needs fewer synaptic changes with a staggered sequence of critical periods, where the critical period of area $n+1$ starts after that of area n has closed, than in a control network where all areas are plastic in parallel. Since our model has no feedback connections, parallel learning of several areas is, at least intuitively, of limited advantage, since synaptic plasticity in area $n+1$ is driven by inputs from an area n that is still undergoing plasticity and thus provides an unstable representation. Yet, going beyond this simple explanation, our results indicate that the usefulness of the representation, as measured by linear classification test accuracy, is even better with critical periods than without. Interestingly, we found that the critical period for area V1 should be shorter than that of later areas assuming that the learning rule and learning rate are identical for all areas.

Even though we believe, based on our experiments with artificial data from the RHM (Fig. 2), that our results are general and applicable outside the domain of vision, we focused our modeling efforts on the ventral visual pathway. Our model of learning in the ventral visual pathway is mostly functional, rather than biophysical. For example, even though our theory highlights differences between apical and basal dendrites, we use a standard artificial neural network with point neurons. This simplification is based on earlier theory work that showed how cross-talk between apical and basal dendrites can lead to a learning rule that matches predictions [40]. Importantly, in our learning rule with lateral predictions, the lateral input into apical dendrites is only used for modulating plasticity but has no direct effect on the firing rates in the feedforward processing pathway. A more detailed simulation of the dendrite [88] or the separation of neurons in a few excitatory and inhibitory cell classes [75] is possible, but goes beyond the scope of the present study.

Compared to earlier studies, the following aspects of our model are important:

(i) In contrast to target-based plasticity rules [38, 40, 46, 51, 61] our model has a multiplicative plasticity rule rather than a subtractive one. In other words, predictive input modulates plasticity directly (e.g., from LTP to LTD) without the need to calculate a difference between prediction and feedforward drive [38, 61]. (ii) In contrast to the non-contrastive plasticity paradigm of [59], we use a contrastive plasticity paradigm in the tradition of earlier work in machine learning [65] -[72]. Interestingly, however, it has recently been shown that contrastive and non-contrastive plasticity paradigms are more intimately linked to each other than previously thought [86]. (iii) In particular, our model is close to the CLAPP model of Illing et al [62], a contrastive plasticity paradigm. However, whereas in [62] the lateral input is an abstract predictive quantity, in our model presented here predictions are transmitted by explicit lateral connections in a spatially arranged cortical area. More precisely, the input of the original CLAPP model never consists of a full image but only of one image patch per time step [62]. Contrastive learning in CLAPP is driven by the distinction between two patches from the same image on one hand (positive sample) versus two patches from two different images on the other hand (contrastive sample) [62]. In the model presented here, however, the network sees the full image. Therefore lateral predictions originate from different locations in the "retinotopic" map, which enables us to interpret the lateral weights as actual connections within a cortical area (Fig. 1).

As a design choice, we assumed that there are no feedback connections. This choice has been made in order to emphasize that feedback connections are, in our model, not *necessary* to achieve representations of input stimuli such that the usefulness of the representation increases from one area to the next. However, the model could be easily extended to include feedback connections so that the loss function for area n influences plasticity within and onto area $n-1$ [30] or more generally that attentional feedback influences learning [89, 90]. Alternatively, the main use of feedback connections present in the brain might not be for learning but rather for defining an attentional focus within the retinotopic map [91] which is particularly important in the presence of cluttered scenes.

4 Methods

The generalized Hebbian learning rule

The network is a VGG-6, which is a convolutional neural network with 6 layers modeling V1 to V2, V4, posterior, central, and inferior temporal cortex respectively. In this method section the term

'layer' therefore refers to one artificial network layer, modeling one cortical area. The convolution kernels are of size 3x3, following measures of the scaling of receptive fields of cells from one area to the next along the visual pathway [92].

The feed-forward processing is given by Equation 1 where ρ is the ReLU non-linearity.

The plasticity in our model relies on the principles of CLAPP, the *Contrastive, Local And Predictive Plasticity* rule [62]. From CLAPP we use the core concept of contrastive/predictive plasticity and layerwise learning , as embodied by the CLAPP Hinge loss where the hinge loss is defined per cortical area [30, 62]. However, we fundamentally change the way visual input is represented so as to move the model from a machine learning setting to a setting that better describes the processing of the ventral visual stream in the cortex.

According to the CLAPP rule, we describe the synaptic weight between presynaptic neuron j and postsynaptic neuron i as described in 2. The first post-synaptic term is simply the derivative of the post-synaptic somatic activity $\text{post}_i(t) = \rho\prime(\sum_j w_{ik}\text{pre}_k)$. The second post-synaptic factor is a predictive term received at the apical dendrites.

Finally, $\eta(t)$ is a layer-specific learning rate that integrates the global, self-supervised learning signal $y \in \{-1, +1\}$ and the layer-specific surprise signal $H \in \{0, +1\}$ that vanishes and suppresses plasticity when the layer is already fit. The surprise signal H emerges from the derivation of the layer-specific loss function and is shared across all neurons in a given layer.

As mentioned above, the learning rule adjusts the connections such that the lateral input predicts the somatic activity [62]. However, the sign of the self-supervised learning signal y switches the direction of plasticity during transitions between *different* images. In this case, the apical predictions, received with the delay Δ, are based on the previous image. These are called "negative samples". Therefore, the sign of plasticity being opposite, the connections are trained such that the apical prediction will highly differ from the somatic activity when saccading from one image to another. Theoretically speaking, this prevents representational collapse. We design our algorithm such that there are as many positive as negative predictions: an image is presented, plasticity happens to match the lateral predictions, then the image is switched and plasticity happens to contrast the activity and prediction.

As mentioned in the Discussion and in contrast to the set-up used in [62], we model the predictive weights w^{lat} as lateral connections in the sense of a retinotopic map. We consider that lateral input to a neuron originates from locations above, below, to the left, and to the right of the postsynaptic neuron, therefore many simultaneous predictions take place in parallel. To model this laterality, we input patches of size 27x27 pixels (instead of 16x16 in [62]) cut out in a grid from the original STL-10 image, and draw comparisons from the representations of almost-neighboring patches (spatially averaged over each patch to drive local consistency) in the four directions. The STL-10 images are randomly cropped to 92x92 pixels.

The Random Hierarchy Model

The Random Hierarchy Model we use is exactly the same as described in [74]; refer to that paper for additional details. The model has n_c classes; each of them can be encoded by any of m synonyms/codewords, each being a different s-tuple ($s = 2$ always) of symbols from a vocabulary of size v, as shown in Fig. 2C. Then, each of these symbols is recursively encoded in the same way. There are L recursive encodings. This means that in the first step an object "cup" can be represented by any AA, AB or CA. Then at the second step, A can be represented by df, fd, or ee; and likewise for B and C with different codewords. Thus each codeword of the object cup has m^s possible encodings after this second step, which means that the object has nm^s possible encodings at the second step of encoding, and $nm^{(s^L)}$ possible codewords after L steps of successive encodings. All our work is based on *full* datasets [74] with $n_c = m = v$, which means that at any step of codeword generation, all possible combinations of symbols are used. There are $P_{max} = m^{(s^L)}$ different codewords or samples in such a dataset. In our Fig 2D, the networks are trained on a training set containing a small multiple of $P^* = n_c m^L$ samples, which is small compared to the full dataset size (for example, for $L = 3$ and $m = 8$ (brown), $P_{max} > 16 \cdot 10^6$ while $P^* = 4096$).

For this non-visual dataset, we do not use our VGG-6 but as in [74], a 1D-convolutional neural network with kernel size and stride adapted to the dataset. The dimensionalities of all hidden layers are kept constant throughout the network and set to a multiple of v^s. Moreover, the activity prediction of our learning rule is computed across two codewords representing the same (or different) object classes (as opposed to two parts of images in the visual case).

Learning goes on until the top-layer loss consistently falls below 0.001 or a maximum number of epochs is reached. The success of learning is tested by linear decoding of object classes from the

representation of one codeword per sample; the test set contains at least 100'000 samples not present in the training set.

For simplification, we use on the RHM datasets simultaneous training rather than developmental training, since they show similar behaviors, and to avoid the question of when to stop training of each layer (which is done heuristically in the visual case).

To show the efficiency of our model on the RHM, our Fig 2E reproduces Fig 3 from [74], using the same color and shape coding.

Critical periods

We model developmental learning by respecting sequential critical periods, *i.e.* training each area only after the previous ones are done learning.

This asynchronicity of the training of different areas enables to train areas to different degrees. Indeed, learning higher areas for a fixed number of epochs while varying the number of epochs for layer 1 shows (Supplementary Figures) that the usefulness (to higher layers) of the representations learned in V1 stops increasing after roughly 20 epochs. No such clear effect was visible for the higher layers. Therefore, all points of Fig 3C are made with the critical period of the first area lasting 20 epochs, while the critical periods of the other areas last equal but varying numbers of epochs. Fig 1B uses the network with best performance in Fig 3C. In Fig 3C, the (total) number of synaptic updates is computed based on the number of neurons, the number of synaptic connections, and the number of epochs used for training each layer; it is represented up to a scaling factor that is similar across our networks.

References

[1] D. Hubel and T. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *Journal of Physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[2] G. Wallis and E. T. Rolls, "Invariant face and object recognition in the visual system.," *Progress in Neurobiology*, vol. 51, pp. 167–194, Feb. 1997.

[3] C. Hung, G. Kreiman, T. Poggio, and J. DiCarlo, "Fast readout of object identity from macaque inferior temporal cortex," *Science*, vol. 310, pp. 863 – 866, 2005.

[4] J. DiCarlo, D. Zoccolan, and N. Rust, "How does the brain solve visual object recognition?," *Neuron*, vol. 73, pp. 415–434, 2012.

[5] D. Hubel and T. Wiesel, "Binocular interaction in striate cortex of kittens reared with artificial squint," *The Journal of Neurophysiology*, vol. 28, pp. 1041–1059, 1965.

[6] D. Feldman, "Dynaptic mechanisms for plasticity in neocortex," *Annual Rev. Neurosci.*, vol. 32, pp. 33–55, 2009.

[7] T. Wiesel and D. Hubel, "Single-cell responses in striate cortex of kittens deprived of vision in one eye.," *J. Neurophysiol.*, vol. 26, pp. 1003–1017, 1963.

[8] T. N. Wiesel, "Postnatal development of the visual cortex and the influence of environment.," *Nature*, vol. 299, pp. 583–591, Oct 1982.

[9] M. Merzenich, R. Nelson, M. Stryker, M. Cynader, A. Schoppmann, and J. Zook, "Somatosensory cortical map changes following digit amputation in adult monkeys," *J. Comparative Neurology*, vol. 224, pp. 591–605, 1984.

[10] N. Li and J. DiCarlo, "Unsupervised natural experience rapidly alters invariant object representation in visual cortex.," *Science*, vol. 321, pp. 1502–1508, 2008.

[11] S. Hofer, T. Mrsic-Flogel, T. Bonhoeffer, and M. Hubener, "Experience leaves a lasting structural trace in cortical circuits.," *Nature*, vol. 457, pp. 313–317, 2009.

[12] M. Insanally, H. Kover, H. Kim, and S. Bao, "Feature-dependent sensitive periods in the development of complex sound representation," *J. Neurosci.*, vol. 29, pp. 5456–5462, 2009.

[13] T. Barkat, D. Polley, and T. Hensch, "A critical period for auditory thalamocortical connectivity," *Nat. Neurosci.*, vol. 14, pp. 1189–1194, 2011.

[14] E. Knudsen, "Sensitive periods in the development of the brain and behavior," *J. Cogn. Neurosci.*, vol. 16, pp. 1412–1425, 2004.

[15] T. Hensch, "Critical period plasticity in local cortical circuits," *Nat. Rev. Neurosci.*, vol. 6, pp. 877–888, 2005.

[16] A. Kral, "Auditory critical periods: a review from system's perspective," *Neurosci.*, vol. 247, pp. 117–133, 2013.

[17] R. Reh, B. Dias, C. Nelson, and T. Hensch, "Critical period regulation across multiple timescales," *Proc. Natl. Acad. Sci. (USA)*, vol. 117, 2020. 23242-23251.

[18] D. Sanes and S. Bao, "Tuning up the developing auditory cns," *Curr. Opin. Neurobiol.*, vol. 19, pp. 188–199, 2009.

[19] T. Hensch, "Critical period regulation," *Ann. Rev. Neurosci.*, vol. 27, pp. 549–579, 2004.

[20] D. O. Hebb, *The Organization of Behavior.* New York: Wiley, 1949.

[21] P. Földiàk, "Learning invariance from transformation sequences," *Neural Computation*, vol. 3, pp. 194–200, 1991.

[22] K. D. Miller, "A model for the development of simple cell receptive fields and the ordered arrangement of orientation columns through activity dependent competition between ON- and OFF-center inputs," *J. Neurosci.*, vol. 14, pp. 409–441, 1994.

[23] K. Körding and P. König, "A learning rule for dynamic recruitment and decorrelation," *Neural Networks*, vol. 13, pp. 1–9, 2000.

[24] B. A. Olshausen and D. J. Field, "How Close Are We to Understanding V1?," *Neural Computation*, vol. 17, pp. 1665–1699, 2005.

[25] R. Miikkulainen, J. A. Bednar, Y. Choe, and J. Sirosh, *Computational maps in the visual cortex.* New York: Springer, 2005.

[26] W. Einhäuser, J. Hipp, J. Eggert, E. Körner, and P. König, "Learning viewpoint invariant object representations using a temporal coherence principle," *Biological Cybernetics*, vol. 93, no. 1, pp. 79–90, 2005.

[27] H. Sprekeler and L. Wiskott, "Analytical derivation of complex cell properties from the slowness principle," in *Proceedings CNS 2006*, 2006.

[28] C. Rozell, D. Johnson, R. Braniuk, and B. Olshausen, "Sparse coding via thresholding and local competition in neural circuits," *Neur. Comput.*, vol. 20, pp. 2526–2563, 2008.

[29] C. S. Brito and W. Gerstner, "Nonlinear hebbian learning as a unifying principle in receptive field formation," *PLOS Comput. Biol.*, vol. 12, p. e1005070, 2016.

[30] A. Marblestone, G. Wayne, and K. Kording, "Toward an integration of deep learning and neuroscience," *Front. Comput. Neurosci.*, vol. 10, p. 94, 2016.

[31] J. Glaser, A. S. Benjamin, R. Farhoodi, and K. Kording, "The roles of supervised machine learning in systems neuroscience," *Progress Neurobiol.*, vol. 175, pp. 126–237, 2019.

[32] S. Khalig-Razavi and N. Kriegeskorte, "Deep supervised, but not unsupervised, models may explain IT cortical representation," *PLOS Comput. Biol.*, vol. 10, p. e1003915, 2014.

[33] D. Yamins and J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nat. Neurosci.*, vol. 19, pp. 356–365, 2016.

[34] B. Richards, T. Lillicrap, and P. B. et al., "A deep learning framework for neuroscience.," *Nat. Neurosci.*, vol. 22, p. 1761–1770, 2019.

[35] M. Schrimpf, J. Kubilius, H. Hong, N. Majaj, ... D.L.K. Yamins, and J. DiCarlo, "Brain-score: Which artificial neural network for object recognition is most brain-like?," *bioRxiv*, 2020. DOI 10.1101/407007.

[36] M. Schrimpf, J. Kubilius, M. Lee, N. Murty, R. Ajemian, , and J. DiCarlo, "Integrative benchmarking to advance neurally mechanistic models of human intelligence," *Neuron*, vol. 108, pp. 413–423, 2020.

[37] F. Crick, "The recent excitement about neural networks," *Nature*, vol. 337, pp. 129–132, 1989.

[38] T. Lillicrap, A. Santoro, L. Marris, C. Akerman, and G. Hinton, "Backpropagation and the brain," *Nat. Rev. Neurosci.*, vol. 21, pp. 335–346, 2020.

[39] T. Lillicrap, D. Cownden, D. Tweed, and C.J.Akerman, "Random synaptic feedback weights support error backpropagation for deep learning," *Nature Communications*, p. 13276, 2016.

[40] R. Urbanczik and W. Senn, "Learning by the dendritic prediction of somatic spiking," *Neuron*, vol. 81, pp. 521–528, 2014.

[41] J. Rombouts, S. Bothe, and R. Roelfsema, "How attention can create synaptic tags for the learning of working memories in sequential tasks," *PLOS Comput. Biol.*, vol. 11, p. e1004060, 2015.

[42] Y. Bengio, B. Scellier, O. Bilaniukl, J. Sacramento, and W. Senn, "Feed- forward initialization for fast inference of deep generative networks is biologically plausible.," *arxiv preprint*, vol. arxiv, p. 1606.01651, 2016.

[43] B. Scellier and Y. Bengio, "Equilibrium propagation: Bridging the gap between energy-based models and backpropagation," *Front. Comput. Neurosci.*, vol. 11, p. 24, 2017.

[44] J. Sacramento, R. Costa, Y. Bengio, and W. Senn, "Dendritic error backpropagation in deep cortical microcircuits," *arxiv preprints*, vol. arxiv, p. 1801:00062, 2017.

[45] J. Whittington and R. Bogacz, "An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity," *Neural Comput.*, vol. 29, pp. 1229–1262, 2017.

[46] J. Sacramento, R. Ponte Costa, Y. Bengio, and W. Senn, "Dendritic cortical microcircuits approximate the backpropagation algorithm," in *Advances in Neural Information Processing Systems 31* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), pp. 8721–8732, Curran Associates, Inc., 2018.

[47] D. Kunin, A. Nayebi, J. Sagastuy-Brena, S. Ganguli, J. Bloom, and D. Yamins, "Two routes to scalable credit assignment without weight symmetry," *PMLR 119: Proc. 37th ICML*, 2020.

[48] I. Pozzi, S. Bothe, and P. Roelfsema, "Attention-gated brain propagation: How the brain can implement reward-based error backpropagation.," in *Adv. Neur. Inform. Proc. Systems 33 (NeurIPS 2020)*, vol. 33, pp. 2516–2526, Curran Associates, Inc., 2020.

[49] Y. Bengio, "Deriving differential target propagation from iterating approximate inverses," *arXiv preprints*, vol. arXiv, p. 2007.15139, 2020.

[50] A. Payeur, J. Guerguiev, F. Zenke, B. Richards, and R. Naud, "Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits," *Nat. Neurosci.*, vol. 24, pp. 1010–1019, 2021.

[51] A. Meulemans, M. Farinha, J. Ordonez, P. Aceituno, J. Sacramento, and B. Grewe, "Credit assignment in neural networks through deep feedback control," *ArXiv preprints*, vol. ArXiv, p. 2106.07887, 2021.

[52] K. Friston, "The free-energy principle: a unified brain theory?," *Nat. Rev. Neurosci.*, vol. 11, pp. 127–138, 2010.

[53] R. Rao and D. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature Neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[54] M. Spratling, "Predictive coding as a model of biased competition in visual attention," *Vision Research*, vol. 48, pp. 1391–1408, 2008.

[55] G. Keller and T. Mrsic-Flogel, "Predictive processing: A canonical cortical computation," *Neuron*, vol. 100, pp. 424–435, 2018.

[56] K. Friston, "A theory of cortical responses," *Phil. Trans. R. Soc. B*, vol. 360, pp. 815–830, 2005.

[57] Y. Huang and R. Rao, "Predictive coding," *Wiley Cognitive Science*, vol. 2, pp. 580–593, 2011.

[58] A. Bastos, W. Usrey, R. Adams, G. Mangun, P. Fries, and K. Friston, "Canonical microcircuits for predictive coding," *Neuron*, vol. 76, pp. 695–711, 2012.

[59] M. Halvagal and F. Zenke, "The combination of hebbian and predictive plasticity learns invariant object representations in deep sensory networks," *Nat. Neurosci.*, vol. 26, pp. 1906–1915, 2023.

[60] L. Williams and A. Holtmaat, "Higher-order thalamocortical inputs gate synaptic long- term potentiation via disinhibition.," *Neuron*, vol. 101, pp. 91–102, 2019.

[61] P. Aceituno, S. de Haan, R. Loidl, and B. Grewe, "Target learning rather than backpropagation explains learning in the mammalian neocortex," *bioRxiv*, vol. preprint, p. DOI 10.1101/2024.04.10.588837, 2024.

[62] B. Illing, J. Ventura, G. Bellec, and W. Gerstner, "Local plasticity rules can learn deep representations using self-supervised contrastive predictions," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 30365–30379, Curran Associates, Inc., 2021.

[63] L. Muckli, F. De Martino, L. Vizioli, L. Petro, F. Swmith, K. Ugurbil, R. Goebel, and E. Yacoub, "Contextual feedback to superficial layers of v1," *Current Biology*, vol. 25, p. 2690–2695, 2015.

[64] A. Keller, M. Roth, and M. Scanziani, "Feedback generates a second receptive field in neurons of the visual cortex," *Nature*, vol. 582, pp. 545–549, 2020.

[65] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, pp. 1771–1800, 2002.

[66] M. Caron, P. Bojanowski, A. Julin, and M. Douze, "Deep clustering for unsupervised learning of visual features.," in *Proc. Europ. Conf. Computer Vision (EECV)*, pp. 132–149, 2018.

[67] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 9726–9735, 2019.

[68] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv*, vol. arXiv, p. 1807.03748, 2019.

[69] S. Lowe, P. O'Connor, and B. Veeling, "Putting an end to end-to-end: Gradient-isolated learning of representations," in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019*, Advances in Neural Information Processing Systems 32 (NeurIPS 2019), 2019. arXiv:1905.11786.

[70] Y. Xiong, M. Ren, and R. Urtasun, "Loco: Local contrastive representation learning.," *Advances inneural information processing systems 34 (NeurIPS2020)*, vol. arXiv, p. 2008.01342, 2020.

[71] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations.," *Proc. of the 37th Int. Conf. Mach. Learn. PMLR*, vol. 119, 2020.

[72] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, pp. 4182–4192, PMLR, 2020.

[73] A. Coates, H. Lee, and A. Ng, "An analysis of single-layer networks in unsupervised feature learnin," *AISTATS and JMLR:W/CP*.

[74] F. Cagnetta, L. Petrini, U. Tomasini, A. Favero, and M. Wyart, "How deep neural networks learn compositional data: The random hierarchy model," *Phys. Rev. X*, vol. 14, p. 031001, 2024.

[75] K. Harris and G. Schepherd, "The neocortical circuit: themes and variations," *Nat. Rev. Neurosci.*, vol. 18, pp. 170–181, 2015.

[76] M. Larkum, "A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex," *Trends Neurosci.*, vol. 36, pp. 141–151, 2013.

[77] P. Sjöström, E. Ranz, A. Roth, and M. Häusser, "Dendritic excitability and synaptic plasticity," *Physiol. Rev.*, vol. 88, pp. 769–840, 2008.

[78] C. Clopath, L. Busing, E. Vasilaki, and W. Gerstner, "Connectivity reflects coding: A model of voltage-based spike-timing-dependent-plasticity with homeostasis.," *Nature Neuroscience*, vol. 13, pp. 344–352, 2010.

[79] H. Markram, J. Sjostrom, and W. Gerstner, "A history of spike-timing dependent plasticity," *Frontiers in Synaptic Neuroscience*, vol. 3, p. 4, 2011.

[80] M. Barry and W. Gerstner, "Fast adaptation to rule switching using neuronal surprise," *PLOS Comput. Biol.*, vol. 20, p. e1011839, 2024.

[81] V. Liakoni, A. Modirshanechi, W. Gerstner, and J. Brea, "Learning in volatile environments with the bayes factor surprise," *Neur. Comput.*, vol. 33, pp. 269–340, 2021.

[82] J. Aru, M. Suzuki, and M. Larkum, "Cellular mechanisms of conscious processing," *Trends Cogn. Sci.*, vol. 24, pp. 814–825, 2020.

[83] E. Kowler, E. Anderson, B. Dosher, and E. Blaser, "The role of attention in the programming of saccades," *Vision Res.*, vol. 35, pp. 1897–1916, 1995.

[84] J. McFarland, A. G. Bondy, R. Saunders, B. Cumming, and D. Butts, "Saccadic modulation of stimulus processing in primary visual cortex.," *Nat. Commun.*, vol. 6, pp. 1–14, 2015.

[85] N. Li and J. DiCarlo, "Neuronal learning of invariant object representation in the ventral visual stream is not dependent on reward.," *J. Neuroscience*, vol. 32, pp. 6611–6620, 2012.

[86] Q. Garrido, Y. Chen, A. Bardes, L. Najman, and Y. Lecun, "On the duality between contrastive and non-contrastive self-supervised learning," *Elenvth Intern. Conf. Learning Repr.*, 2023. DOI 10.48550/arXiv.2206.02574.

[87] A. Patel, T. Nguyen, and R. Baraniuk, "A probabilistic theory of deep learning," *arXiv preprints*, vol. arXiv, p. 1504.00641, 2015.

[88] P. Poirazi and B. W. Mel, "Impact of active dendrites and structural plasticity on the memory capacity of neural tissue.," *Neuron*, vol. 29, no. 3, pp. 779–796, 2001.

[89] P. Roelfsema and A. van Ooyen, "Attention-gated reinforcement learning of internal representations for classification," *Neural Computation*, vol. 17, pp. 2176–2214, 2005.

[90] P. Roelfsema, A. van Ooyen, and T. Watanabe, "Perceptual learning rules based on reinforcers and attention," *Trends in Cogn. Sci.*, vol. 14, pp. 64–71, 2010.

[91] V. Lamme and P. Roelfsema, "The distinct modes of vision offered by feedforward and recurrent processing," *Trends in neurosci.*, vol. 23, pp. 571–579, 2000.

[92] H. R. Wilson and F. Wilkinson, "From orientations to objects: Configural processing in the ventral stream," *Journal of Vision*, vol. 15, p. 4, May 2015.