Title

Deep Predictive Learning in Neocortex and Pulvinar

Journal

Journal of Cognitive Neuroscience, 33(6)

ISSN

0898-929X

Authors

O'Reilly, Randall C
Russin, Jacob L
Zolfaghar, Maryam
et al.

Publication Date

2021-05-01

DOI

10.1162/jocn_a_01708

Peer reviewed

Published in final edited form as:

J Cogn Neurosci. 2021 May 01; 33(6): 1158–1196. doi:10.1162/jocn_a_01708.

Deep Predictive Learning in Neocortex and Pulvinar

Randall C. O'Reilly,

Jacob L. Russin,

Maryam Zolfaghar,

John Rohrlich

Department of Psychology, Computer Science, and Center for Neuroscience, University of California Davis, 1544 Newton Ct, Davis, CA 95618

Abstract

How do humans learn from raw sensory experience? Throughout life, but most obviously in infancy, we learn without explicit instruction. We propose a detailed biological mechanism for the widely-embraced idea that learning is driven by the differences between predictions and actual outcomes (i.e., *predictive error-driven learning*). Specifically, numerous weak projections into the pulvinar nucleus of the thalamus generate top-down predictions, and sparse *driver* inputs from lower areas supply the actual outcome, originating in layer 5 intrinsic bursting (5IB) neurons. Thus, the outcome representation is only briefly activated, roughly every 100 ms (i.e., 10 Hz, *alpha*), resulting in a *temporal difference error signal*, which drives local synaptic changes throughout the neocortex. This results in a biologically-plausible form of error backpropagation learning. We implemented these mechanisms in a large-scale model of the visual system, and found that the simulated inferotemporal (IT) pathway learns to systematically categorize 3D objects according to invariant shape properties, based solely on predictive learning from raw visual inputs. These categories match human judgments on the same stimuli, and are consistent with neural representations in IT cortex in primates.

The fundamental epistemological conundrum of how knowledge emerges from raw experience has challenged philosophers and scientists for centuries. Although there have been significant advances in cognitive and computational models of learning (Ashby & Maddox, 2011; LeCun, Bengio, & Hinton, 2015; Watanabe & Sasaki, 2015) and in our understanding of the detailed biochemical basis of synaptic plasticity (Cooper & Bear, 2012; Lüscher & Malenka, 2012; Shouval, Bear, & Cooper, 2002; Urakubo, Honda, Froemke, & Kuroda, 2008), there is still no widely-accepted answer to this puzzle that is clearly supported by known biological mechanisms and also produces effective learning at the computational and cognitive levels. The idea that we learn via an active *predictive* process was advanced by Helmholtz in his *recognition by synthesis* proposal (von Helmholtz, 1867), and has been widely embraced in a range of different frameworks (Clark, 2013; Dayan, Hinton, Neal, & Zemel, 1995; de Lange, Heilbron, & Kok, 2018; J. Elman et al., 1996; J. L. Elman, 1990; Friston, 2005; George & Hawkins, 2009; Hawkins & Blakeslee, 2004;

---

oreilly@ucdavis.edu .

Kawato, Hayakawa, & Inui, 1993; Mumford, 1992; Rao & Ballard, 1999; Summerfield & de Lange, 2014).

Here, we propose a detailed biological mechanism for a specific form of *predictive error-driven learning* based on distinctive patterns of connectivity between the neocortex and the higher-order nuclei of the thalamus (i.e., the pulvinar) (S. M. Sherman & Guillery, 2006; Usrey & Sherman, 2018). We hypothesize that learning is driven by the difference between top-down predictions, generated by numerous weak projections into the thalamic relay cells (TRCs) in the pulvinar, and the actual outcomes supplied by sparse, strong *driver* inputs from lower areas. Because these driver inputs originate in layer 5 intrinsic bursting (5IB) neurons, the outcome is only briefly activated, roughly every 100 ms (i.e., 10 Hz, *alpha*). Thus, the prediction error is a *temporal difference* in activation states over the pulvinar, from an earlier prediction to a subsequent burst of outcome. This temporal difference can drive local synaptic changes throughout the neocortex, supporting a biologically-plausible form of error backpropagation that improves the predictions over time (Ackley, Hinton, & Sejnowski, 1985; Bengio, Mesnard, Fischer, Zhang, & Wu, 2017; Hinton & McClelland, 1988; Lillicrap, Santoro, Marris, Akerman, & Hinton, 2020; O'Reilly, 1996; Whittington & Bogacz, 2019). The temporal-difference form of error-driven learning contrasts with prevalent alternative hypotheses that require a separate population of neurons to compute a prediction error *explicitly* and transmit it directly through neural firing (Friston, 2005, 2010; Kawato et al., 1993; Lotter, Kreiman, & Cox, 2016; Ouden, Kok, & Lange, 2012; Rao & Ballard, 1999).

In the following, our primary objective is to describe the hypothesized biologically based mechanism for predictive error-driven learning, contrast it with other existing proposals regarding the functions of this thalamocortical circuitry and other ways that the brain might support predictive learning, and evaluate it relative to a wide range of existing anatomical and electrophysiological data. We provide a number of specific empirical predictions that follow from this functional view of the thalamocortical circuit, which could potentially be tested by current neuroscientific methods. Thus, this work proposes a clear functional interpretation of this distinctive thalamocortical circuitry that contrasts with existing ideas in testable ways.

A second major objective is to implement this predictive error-driven learning mechanism in a large-scale computational model that faithfully captures its essential biological features, to test whether the proposed learning mechanism can drive the formation of cognitively useful representations. In particular, we ask a critical question for any predictive-learning model: can it develop high-level, abstract representations while learning from nothing but predicting low-level visual inputs. Most visual object recognition models that provide a reasonable fit to neurophysiological data rely on large human-labeled datasets to explicitly train abstract category information via error-backpropagation (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Rajalingham et al., 2018). Thus, it is perhaps not too surprising that the higher layers of these models, which are closer to these category output labels, exhibited a greater degree of categorical organization.

Through large-scale simulations based on the known structure of the visual system, we found that our biologically based predictive learning mechanism developed high-level, abstract representations that signif icantly diverge from the similarity structure present in the lower layers of the network, and systematically categorize 3D objects according to invariant shape properties. Furthermore, we found in an experiment using the same stimuli that these categories match human similarity judgments, and that they are also qualitatively consistent with neural representations in inferotemporal (IT) cortex in primates (Cadieu et al., 2014). In addition, we show that comparison predictive backpropagation models lacking these biological features (Lotter et al., 2016) did not learn object categories that go beyond the visual input structure. Thus, there may be some important features of the biologically based model that enable this ability to learn higher-level structure beyond that of the raw inputs.

It is important to emphasize that our objectives for these simulations are *not* to produce a better machine-learning (ML) algorithm *per se*, but rather to test whether our biologically based model can capture some of the known high-level, cognitive phenomena that the mammalian brain learns. Thus, we explicitly dissuade readers from the inevitable desire to evaluate the importance of our model based on differences in narrow, performance-based ML metrics. As discussed later, there are various engineering-level issues regarding the biologically based model's computational cost and performance, that currently limit its ability to compete with simpler, much larger-scale backpropagation models, but we do not think these are relevant to the evaluation of the scientific questions of relevance here. In short, this model is an instantiation of a scientific theory, and it should be evaluated on its ability to explain a wide range of data across multiple levels of analysis, just as every other scientific theory is evaluated.

The remainder of the paper is organized as follows. First, we provide a concise overview of the biologically based predictive error-driven learning framework, including the most relevant neural data. Then, we present a small-scale implementation of the model that learns a probabilistic grammar, to illustrate the basic computational mechanisms of the theory. This is followed by the large-scale model of the visual system, which learns by predicting over brief movies of 3D objects rotating and translating in space. We evaluate this model and compare it to two other predictive learning models that directly use error-backpropagation, based on current deep convolutional neural network (DCNN) mechanisms. Then, we circle back to discuss the relevant biological data in greater detail, along with testable predictions that can differentiate this account from other existing ideas. Finally, we conclude with a discussion of related models and outstanding issues.

Predictive Error-driven Learning in the Neocortex and Pulvinar

Figure 1 shows the thalamocortical circuits characterized by S. M. Sherman and Guillery (2006) (see also S. M. Sherman & Guillery, 2013; Usrey & Sherman, 2018), which have two distinct projections converging on the principal thalamic relay cells (TRCs) of the *pulvinar*, the primary thalamic nucleus that is interconnected with higher-level posterior cortical visual areas (Arcaro, Pinsk, & Kastner, 2015; Halassa & Kastner, 2017; Shipp, 2003). One projection consists of numerous, weaker connections originating in deep layer

VI of the neocortex (the 6CT corticothalamic projecting cells), which we hypothesize generate a top-down prediction on the pulvinar. The other is a sparse (Rockland, 1996, 1998) and strong *driver* pathway that originates from lower-level layer 5 intrinsic bursting cells (5IB), which we hypothesize provide the outcome. These 5IB neurons fire discrete bursts with intrinsic dynamics having a period of roughly 100 ms between bursts (Connors, Gutnick, & Prince, 1982; Franceschetti et al., 1995; Larkum, Zhu, & Sakmann, 1999; Saalmann, Pinsk, Wang, Li, & Kastner, 2012; Silva, Amitai, & Connors, 1991), which is thought to drive the widely-studied *alpha* frequency of ~ 10 Hz that originates in cortical deep layers and has important effects on a wide range of perceptual and attentional tasks (Buffalo, Fries, Landman, Buschman, & Desimone, 2011; Clayton, Yeung, & Kadosh, 2018; Jensen, Bonnefond, & VanRullen, 2012; K. Mathewson, Gratton, Fabiani, Beck, & Ro, 2009; VanRullen & Koch, 2003). Critically, unlike many other such bursting phenomena, this 5IB bursting occurs in awake animals (Luczak, Bartho, & Harris, 2009, 2013; Sakata & Harris, 2009, 2012), consistent with presence of alpha in awake behaving states.

The existing literature generally characterizes the 6CT projection as *modulatory* (S. M. Sherman & Guillery, 2013; Usrey & Sherman, 2018), but a number of electrophysiological recordings from awake, behaving animals clearly show sustained, continuous patterns of neural firing in pulvinar TRC neurons, which is not consistent with the idea that they are only being driven by their phasic bursting 5IB inputs (Bender, 1982; Bender & Youakim, 2001; Komura, Nikkuni, Hirashima, Uetake, & Miyamoto, 2013; Petersen, Robinson, & Keys, 1985; Robinson, 1993; Saalmann et al., 2012; Zhou, Schafer, & Desimone, 2016). Indeed, these recordings show that pulvinar neural firing generally resembles that of the visual areas with which they interconnect, in terms of neural receptive field properties, tuning curves, etc. This is important because our predictive learning framework requires that these 6CT top-down projections be capable of directly driving TRC activity. Specifically, in contrast to the standard view, the core idea behind our theory is that the top-down 6CT projections drive a *predicted* activity pattern across the extent of the pulvinar, which precedes the subsequent *outcome* activation state driven by the strong 5IB inputs.

Figure 2 illustrates the temporal evolution of activity states according to our predictive learning theory, which is somewhat challenging to convey because the critical signals driving learning unfold *over time* (Kachergis, Wyatte, O'Reilly, de Kleijn, & Hommel, 2014; O'Reilly, Wyatte, & Rohrlich, 2014, 2017). We hypothesize that synaptic plasticity throughout the cortex is sensitive to the resulting *temporal differences* that emerge initially in the pulvinar. Thus, unlike other models (as we discuss in depth later) the prediction error here is not captured directly in the firing of a special population of error-coding neurons, but rather remains as a temporal difference error signal.

The figure shows a single 125 ms time window of a 100 ms alpha cycle for the purposes of illustration (the actual timing is likely to be more dynamic as discussed next). The activity state in pulvinar TRC neurons, representing a prediction, as driven by the top-down 6CT projections, should develop during the first ~ 75 ms, when the 5IB neurons are paused between bursting. Then the final ~ 25 ms largely reflects the strong 5IB bottom-up ground-truth driver inputs when they burst. Thus, the prediction error signal is reflected in the temporal difference of these activation states as they develop over time. In other words,

our hypothesis is that the pulvinar is directly representing either the top-down prediction or the bottom-up outcome at any given time, and the temporal difference between these states implicitly encodes a prediction error. While the deep 6CT layer is involved in generating a top-down prediction over the pulvinar, the superficial layer neurons continuously represent the current state, simultaneously incorporating bottom-up and top-down constraints via their own connections with other areas. To ensure that the prediction is not directly influenced by this current state representation (i.e., "peeking at the right answer"), it is important that the 6CT neurons encode temporally delayed information, consistent with available data (Harris & Shepherd, 2015; Sakata & Harris, 2009; Thomson, 2010).

The actual biological system is likely to be much more dynamic than the simplistic cartoon with rigid 100 ms timing as shown in Figure 2, based on a set of neural mechanisms that can work together to enable it to more flexibly entrain the predictive learning cycle to the environment. These mechanisms would also tend to increase activity and learning associated with unexpected outcomes relative to expected ones, consistent with the observed *expectation suppression* phenomena (Bastos et al., 2012; Meyer & Olson, 2011; Summerfield, Trittschuh, Monti, Mesulam, & Egner, 2008; Todorovic, van Ede, Maris, & de Lange, 2011).

Specifically, various underlying mechanisms result in neural *adaptation*, which is generally thought to increase neural activity and learning associated with novel inputs relative to recently familiar ones (Abbott, Varela, Sen, & Nelson, 1997; Brette & Gerstner, 2005; Grill-Spector, Henson, & Martin, 2006; Hennig, 2013; Müller, Metha, Krauskopf, & Lennie, 1999). In the case where outcomes are consistent with prior predictions (i.e., the predictions are accurate), the same population of neurons across pulvinar and cortex should be active over time, whereas unpredicted outcomes will generally activate new subsets of neurons in superficial cortical layers representing the current state. Thus, due to adaptation, there should be a phasic increase in activity in these superficial neurons at the onset of unpredicted stimuli relative to predicted ones. Furthermore, the 5IB neurons downstream of these superficial neurons may be particularly responsive to these phasic activity increases, causing their bursting to coincide preferentially with unexpected outcomes, thereby driving the phase resetting of the alpha cycle to such events. Thus, during a sequence of predicted states, the pulvinar may experience relatively weaker or even absent 5IB driving inputs, until an unpredicted stimulus arises. At this point, error-driven learning would be more strongly engaged as a function of the phasic release from adaptation and 5IB burst activation. We discuss these dynamics more later in the context of the comparison with explicit error coding models.

We also hypothesize that 5IB bursting preferentially drives the synaptic plasticity processes to take place at that time, due the strong driving nature of the outputs from these neurons. In computational terms originating with the *Boltzmann Machine* (Ackley et al., 1985; Hinton & Salakhutdinov, 2006), this anchors the target or *plus* phase to be at this point of 5IB bursting. Furthermore, this means that the predictive nature of the prior minus phase naturally emerges just by virtue of it being the state prior to 5IB bursting: the learning rule automatically causes that prior state to better anticipate the subsequent state. Thus, even if no prediction was initially generated, learning over multiple iterations will work to

create one, to the extent that a reliable prediction can be generated based on internal states and environmental inputs. Likewise, assuming relevant activity traces naturally persist over timescales longer than the alpha cycle, this predictive learning process can take advantage of any such remaining traces to learn across these longer timescales, even though it is operating at the faster alpha scale.

In short, learning always happens whenever something unexpected occurs, at any point, and drives the development of predictions immediately prior, to the extent such predictions are possible to generate. In the typical lab experiment where phasic stimuli are presented without any predictable temporal sequence (which is uncharacteristic of the natural world), there may often be no significant prediction prior to stimulus onset, and we would expect such stimuli to reliably drive 5IB bursting, which is consistent with available electrophysiological data (Bender, 1982; Bender & Youakim, 2001; Komura et al., 2013; Luczak et al., 2009, 2013; Petersen et al., 1985; Robinson, 1993; Zhou et al., 2016). Thus, unlike Figure 2, such situations would *start* with a 5IB-triggered plus phase, without a significant minus phase prior to that.

As may be evident by this point, we are mainly focused on *prediction* in the sense of the humorous quote: "prediction is very difficult, especially about the future" (attributable to Danish author Robert Storm Petersen), whereas this term is potentially confusingly used in a much broader sense in most Bayesian-inspired predictive coding frameworks (de Lange et al., 2018; Friston, 2005; Rao & Ballard, 1999). These frameworks use "prediction" to encompass everything from genetic biases to the results of learning in the feedforward synaptic pathways to top-down filling-in or biasing of the current stimulus properties, and fairly rarely use it in the "about the future" sense. We think these different phenomena are each associated with different neural mechanisms at different time scales (O'Reilly, Hazy, & Herd, 2016; O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012; O'Reilly, Wyatte, Herd, Mingus, & Jilk, 2013), and thus prefer to treat them separately, while also recognizing that they can clearly interact as well.

Thus, our use of the term *prediction* here refers specifically to *anticipatory* neural firing that predicts subsequent stimuli. We use the term *postdiction* to refer to the operation of this predictive mechanism after a stimulus has been initially processed (to consolidate and more deeply encode, as in an auto-encoder model), and distinguish both from *top-down excitatory biasing*, which directly influences the online superficial layer neural representations of the current stimulus (Desimone & Duncan, 1995; E. K. Miller & Cohen, 2001; O'Reilly et al., 2013; Reynolds, Chelazzi, & Desimone, 1999). Finally, many discussions of prediction error in the literature include late, frontally-associated processes such as those associated with the P300 ERP component (Holroyd & Coles, 2002). We specifically exclude these from the scope of the mechanisms described here, which are anticipatory, fast, and low-level, as is appropriate for the posterior cortical sensory processing areas that interconnect with the pulvinar.

Computational Properties of Predictive Learning in the Thalamocortical Circuits

We next elaborate the connections between the computational properties required for predictive learning, and the properties of the circuits interconnecting cortex and the pulvinar,

which appear to be notably well suited for their hypothesized role in predictive learning. We begin with a relatively established interpretation of superficial layer processing, to contextualize subsequent points about the special functions required of the deep layers and the thalamus.

- **The superficial cortical layers continuously represent the current state:** The superficial layer pyramidal neurons are densely and bidirectionally interconnected with other cortical areas, and update quickly to new stimulus inputs, with continuous, relatively rapid firing (i.e., up to about 100 Hz for preferred stimuli). These neurons integrate higher-level top-down information with bottom-up sensory information to resolve ambiguities, focus attention, fill in missing information, and generally enhance the consistency and quality of the online representations (Desimone & Duncan, 1995; Hopfield, 1984; E. K. Miller & Cohen, 2001; O'Reilly et al., 2016, 2012, 2013; Reynolds et al., 1999; Rumelhart & McClelland, 1982). As noted above, we distinguish this form of top-down processing, which is often most evident during the period *after* stimulus onset (Lee & Mumford, 2003), from the specifically predictive, anticipatory sort.

- **Predictions must be insulated against receiving current state information (it isn't prediction if you already know what happens):** Given that the superficial layers are continuously updating and representing the current state, some kind of separate neural system insulated from this current state information must be used to generate predictions, otherwise the prediction system can just "cheat" and directly report the current state. It may seem counter-intuitive, but making the prediction task *harder* is actually beneficial, because that pushes the learning to capture deeper, more systematic regularities about how the environment evolves over time. In other words, like any kind of cheating, the cheater itself is cheated because of the reduced pressure to learn, and learning is the real goal.

- **Predictions take time and space to generate:** Non-trivial predictions likely require the integration of multiple converging inputs from a range of higher-level cortical areas, each encoding different dimensions of relevance (e.g., location, motion, color, texture, shape, etc). Thus, sufficient time and space (i.e., neural substrates with relevant connectivity) must be available to integrate these signals into a coherent predicted state, and per the above point, these substrates must be separated from the influence of current state information. This fits with the properties of the layer 6CT neurons and their deep layer inputs, which we hypothesize are insulated from superficial-layer firing by virtue of being driven locally by the 5IB bursting within their own cortical microcolumn, such that the inter-bursting pause period provides a time window when these deep layers can integrate and generate the prediction.

  Biologically, this is consistent with the delayed responses of 6CT neurons (Harris & Shepherd, 2015; Sakata & Harris, 2009; Thomson, 2010). Computationally, these neurons function much like the simple recurrent network (SRN) context layer updating (J. L. Elman, 1990; Jordan, 1989) which reflects the prior trial's

state, as discussed in detail in the Appendix. The overall duration of the alpha cycle may represent a reasonable compromise between the prediction integration time and the need to keep up with predictions tracking changes in the world. Notably, films are typically shown at just over 2 times the alpha frequency (24 Hz), suggesting a Nyquist sampling relative to the underlying alpha processing.

- **The predicted state must be directly aligned with the outcome state it predicts:** A prediction error is a difference between two states, so these prediction and outcome states must be directly comparable such that their difference meaningfully represents the actual prediction error, and not some other kind of irrelevant encoding differences. In other words, the prediction and the outcome must be represented in the same "language", so that the "words" from the prediction can be directly compared against those of the outcome — if the prediction was in Japanese and the outcome in English, it would be hard to tell whether the prediction was correct or not! Thus, a common neural substrate with two different input pathways is required, one reflecting the prediction and the other the outcome, so that both converge onto the same representational system within this common neural substrate. This fits well with the two pathways converging into the pulvinar: the 6CT top-down prediction-generation pathway and the lower-level 5IB driving inputs.
- **The outcome signal should be as *veridical* as possible (i.e., directly reflecting the bottom-up outcome), and should arise from lower areas in the hierarchy relative to the corresponding predictive 6CT inputs:** Given that the outcome is the driver of learning, if it were to be corrupted or inaccurate, then everything that is learned would then be suspect. To the extent that delusional thinking is present in all people (some moreso than others perhaps) this principle must be violated at some level, but for the lowest levels of the perceptual system at least, it is important that strongly grounded, accurate training signals drive learning. The bottom-up, sparse, strongly driving nature of the 5IB projections to the pulvinar can directly convey such veridical outcome signals, and ensure that they dominate the activation of their TRC targets. Based on indirect available data, it is likely that each pulvinar TRC neuron receives only roughly 1–6 driver inputs (S. M. Sherman & Guillery, 2006, 2011), such that these sparse inputs directly convey the signal from lower layers, without much further mixing or integration (which could distort the nature of the signal). Furthermore, these inputs are likely not plastic (Usrey & Sherman, 2018), again consistent with a need for unaltered, veridical signals. Lastly, the TRC neurons are distinctive in having no significant lateral interconnectivity (S. M. Sherman & Guillery, 2006), enabling them to faithfully represent their inputs. These properties led Mumford (1991) to characterize the pulvinar as a *blackboard*, and we further suggest the metaphor of a *projection screen* upon which the predictions are projected.
- **The prediction error must drive learning to reduce subsequent prediction errors:** Obviously, this is the goal of prediction error learning in the first place, and given that the cortex is what generates predictions, it must be capable of learning based on prediction error signals represented over the pulvinar.

Computationally, the critical problem here is *credit assignment*: how do the error signals direct learning in the proper direction for each individual neuron, to reduce the overall prediction error? The error backpropagation procedure solves this problem (Rumelhart, Hinton, & Williams, 1986), but requires biologically implausible retrograde signaling across the entire network of neural communication (Crick, 1989), to propagate the error proportionally back along the same channels that drive forward activation. Bidirectional connections, which are ubiquitous in the cortex (Felleman & Van Essen, 1991; Markov, Ercsey-Ravasz, et al., 2014) and computationally beneficial for other reasons as noted earlier, can eliminate that problem by "implicitly" propagating error signals via standard neural communication mechanisms along both directions of connectivity (O'Reilly, 1996).

This solution to the credit assignment problem relies on a *temporal difference* error signal, as originally developed for the *Boltzmann machine* (Ackley et al., 1985). The bidirectional neural communication at one point in time is encoding and sharing the prediction among the entire network of neurons. Then, this same network of connections is reused at another point in time to encode and communicate the outcome. Mathematically, the difference in activation state across these two points in time, locally at each individual neuron, provides an accurate estimate of the error backpropagation gradient (O'Reilly, 1996). In effect, this temporal difference tells each neuron which direction it needs to change its activation state to reduce the overall error. The reuse of the very same network of connections across both points in time ensures the overall alignment of the two activation states, as noted above, such that this temporal difference precisely represents the error signal. While various other schemes for error-driven learning in biologically-plausible networks have been proposed (e.g., Bengio et al., 2017; Lillicrap et al., 2020; Whittington & Bogacz, 2019), the temporal-difference framework with bidirectional connectivity provides a particularly good fit with the natural temporal ordering of predictive learning (prediction then outcome) and the extensive bidirectional connectivity of the thalamocortical circuits (Shipp, 2003).

- **Temporal differences in activation state across the alpha cycle, between prediction and outcome states, must drive synaptic plasticity:** The final step needed to connect all of the elements above is that neurons actually modify their synaptic strengths in proportion to the temporal-difference error signal. We have recently provided a fully explicit mechanism for this form of learning (O'Reilly et al., 2012), based on a biologically-detailed model of spike timing dependent plasticity (STDP) (Urakubo et al., 2008). We showed that when activated by realistic Poisson spike trains, this STDP model produces a nonmonotonic learning curve similar to that of the BCM model (Bienenstock, Cooper, & Munro, 1982), which results from competing calcium-driven postsynaptic plasticity pathways (Cooper & Bear, 2012; Shouval et al., 2002). As in the BCM framework, we hypothesized that the threshold crossover point in this nonmonotonic curve moves dynamically — if this happens on the

alpha timescale (Lim et al., 2015), then it can reflect the prediction phase of activity, producing a net error-driven learning rule based on a subsequent calcium signal reflecting the outcome state. The resulting learning mechanism naturally supports a combination of both BCM-style hebbian learning and error-driven learning, where the BCM component acts as a kind of regularizer or bias, similar to weight decay (O'Reilly & Munakata, 2000; O'Reilly et al., 2012).

Thus, remarkably, the pulvinar and associated thalamocortical circuitry appears to provide *precisely* the necessary ingredients to support predictive error-driven learning, according to the above analysis. Interestingly, although S. M. Sherman and Guillery (2006) did not propose a predictive learning mechanism as just described, they did speculate about a potential role for this circuit in motor forward-model learning and the predictive remapping phenomenon (S. M. Sherman & Guillery, 2011; Usrey & Sherman, 2018). In addition, Pennartz, Dora, Muckli, and Lorteije (2019) also suggested that the pulvinar may be involved in predictive learning, but within the explicit error-coding framework and not involving the detailed aspects of the above-described circuitry.

It bears emphasizing the synergy between the various considerations above for the benefits of the pause in 5IB firing between bursts. First, this pause is critical for creating the time window when the predictive network is representing and communicating the prediction state, without influence from the outcome state. Further, it creates the temporal difference in activation state in the pulvinar between prediction and outcome, which is needed for driving error-driven learning. Thus, for both the 6CT and pulvinar layers, the periodic pausing of 5IB neurons is essential for creating the predictive learning dynamic. Interestingly, by these principles, the lack of such burst / pause dynamics in the driver inputs to first-order sensory thalamus areas such as the LGN and MGN (S. M. Sherman & Guillery, 2006) means that these areas should *not* be directly capable of error-driven predictive learning. This is consistent with a number of models and theoretical proposals suggesting that primary sensory areas may learn predominantly through hebbian-style self-organizing mechanisms (Bednar, 2012; K. D. Miller, 1994). Nevertheless, primary sensory areas do receive "collateral" error signals from the pulvinar (Shipp, 2003), which could provide some useful indirect error-driven learning signals.

Note that this form of temporal-difference learning signal is distinct from the widely-used TD (temporal-difference) model in reinforcement learning (Sutton & Barto, 1998), which is scalar, and applies to reward expectations, not sensory predictions (although see Gardner, Schoenbaum, & Gershman, 2018 and Dayan, 1993 for potential connections between these two forms of prediction error). Finally, as we discuss later, this proposed predictive role for the pulvinar is compatible with the more widely-discussed role it may play in attention (Bender & Youakim, 2001; Fiebelkorn & Kastner, 2019; LaBerge & Buchsbaum, 1990; Saalmann & Kastner, 2011; Snow, Allen, Rafal, & Humphreys, 2009; Zhou et al., 2016). Indeed, we think these two functions are synergistic (i.e., you predict what you attend, and vice-versa; Richter & de Lange, 2019), and have initial computational results consistent with this idea.

Predictive Learning of Temporal Structure in a Probabilistic Grammar

To illustrate and test the predictive learning abilities of this biologically based model, we first ran a classical test of sequence learning (Cleeremans & McClelland, 1991; Reber, 1967) that has been explored using simple recurrent networks (SRNs) (J. L. Elman, 1990; Jordan, 1989). The biologically based model was implemented using the *Leabra* algorithm, which is a comprehensive framework that uses conductance-based point neuron equations, inhibitory competition, bidirectional connectivity, and the biologically plausible temporal difference learning mechanism described above (O'Reilly, 1996, 1998; O'Reilly et al., 2016; O'Reilly & Munakata, 2000; O'Reilly et al., 2012). Leabra serves as a model of the bidirectionally connected processing in the cortical superficial layers, and has been used to simulate a large number of different cognitive neuroscience phenomena. It is described in the Appendix, which also provides a detailed mapping between the SRN and our biological model.

As shown in Figure 3, sequences were generated according to a finite state automaton (FSA) grammar, as used in implicit sequence learning experiments by Reber (1967). Each node has a 50% random branching to two different other nodes, and the labels generated by node transitions are locally ambiguous (except for the B=begin and E=end states). Thus, integration over time and across many iterations are required to infer the systematic underlying grammar. It is a reasonably challenging task for SRNs and people to learn and provides an important validation of the power of these predictive learning mechanisms. Given the random branching, accurately predicting the specific path taken is impossible, but we can score the model's output as correct if it activates either or both of the possible branches for each state.

The model (Figure 4) required around 20 epochs of 25 sequences through the grammar to learn it to the point of making no prediction errors for 5 epochs in a row (which guarantees that it had completely learned the task). This model is available in the standard *emergent* distribution, at https://github.com/emer/leabra/tree/master/examples/deepfsa. A few steps through a sequence are shown in the figure, illustrating how the CT context layer, which drives the P pulvinar layer prediction, represents the information present on the *previous* alpha cycle time step. Thus, the network is attempting to predict the current Input state, which then drives the pulvinar plus phase at the end of each alpha cycle, as shown in the last panel. On each trial, the difference between plus and minus phases locally over each cortical neuron drives its synaptic weight changes, which accumulate over trials to allow accurate prediction of the sequences, to the extent possible given their probabilistic nature.

Predictive Learning of Object Categories in IT Cortex

Now we describe a large-scale, systems-neuroscience implementation of the proposed thalamocortical predictive error-driven learning framework, in a model of visual predictive learning (Figure 5). Our second major objective, and a critical question for predictive learning, is determining whether the model can develop high-level, abstract ways of representing the raw sensory inputs, while learning from nothing but predicting these low-level visual inputs. We showed the model brief movies of 156 3D object exemplars drawn from 20 different basic-level categories (e.g., car, stapler, table lamp, traffic cone,

etc.) selected for their overall shape diversity from the CU3D-100 dataset (O'Reilly et al., 2013). The objects moved and rotated in 3D space over 8 movie frames, where each frame was sampled at the alpha frequency (Figure 5b). Because the motion and rotation parameters were generated at random on each sequence, this dataset consists of 512,000 unique images, and there is no low-dimensional object category training signal, so the usual concerns about overfitting and training vs. testing sets are not applicable: our main question is what kind of representations self-organize as result of this purely visual experience.

There were also saccadic eye movements every other frame, introducing an additional, realistic, predictive-learning challenge. An efferent copy signal enabled full prediction of the effects of the eye movement, and allows the model to capture the signature predictive remapping phenomenon (Cavanagh, Hunt, Afraz, & Rolfs, 2010; Duhamel, Colby, & Goldberg, 1992; Neupane, Guitton, & Pack, 2017). The *only* learning signal available to the model was the prediction error generated by the temporal difference between what it predicted to see in the V1 input in the next frame and what was actually seen.

As described in detail in the Appendix, our model was constructed to capture critical features of the visual system, including the major division between a dorsal *Where* and ventral *What* pathway (Ungerleider & Mishkin, 1982), and the overall hierarchical organization of these pathways derived from detailed connectivity analyses (Felleman & Van Essen, 1991; Markov, Ercsey-Ravasz, et al., 2014; Markov, Vezoli, et al., 2014; Rockland & Pandya, 1979). In addition to these biological constraints, we conducted extensive exploration of the connectivity and architecture space, and found a remarkable convergence between what worked functionally and the known properties of these pathways (O'Reilly et al., 2017). For example, the feedforward pathway has projections from lower-level superficial layers to superficial layers of higher levels, while feedback originated in both the superficial and deep and projected back to both (Felleman & Van Essen, 1991; Rockland & Pandya, 1979). Also, consistent with the core features of the pulvinar pathways discussed above, deep layer predictive (6CT) inputs originated in higher levels, while driver (5IB) inputs originated in lower levels. For simplicity we organized the model layers in terms of these driver inputs, whereas the topographic organization of pulvinar in the brain is organized more according to the 6CT projection loops (Shipp, 2003).

Another important set of parameters are the strength of deep-layer recurrent projections, which influence the timescale of temporal integration, producing a simple biologically based version of *slow feature analysis* (Foldiak, 1991; Wiskott & Sejnowski, 2002). We followed the biological data suggesting that recurrence increases progressively up the visual hierarchy (Chaudhuri, Knoblauch, Gariel, Kennedy, & Wang, 2015). It was essential that the *Where* pathway learn first, consistent with extant data (Bourne & Rosa, 2006; Kiorpes, Price, Hall-Haro, & Anthony Movshon, 2012), including early pathways interconnecting LIP and pulvinar (Bridge, Leopold, & Bourne, 2016), and a rare asymmetric pathway, from V1 to LIP (Markov, Ercsey-Ravasz, et al., 2014), providing a direct short-cut for high-level spatial representations in LIP. Results from various informative model architecture and parameter manipulations are discussed below after the primary results from the standard intact model.

Learning curves and other model details are shown in the Appendix. We have also implemented a full *de-novo* replication of the model in a new modeling framework, which also replicated the results shown here. Furthermore, much of the model was originally developed in the context of a set of object-like patterns generated systematically from a set of simple line features O'Reilly et al. (2017), and the parameters that work best in terms of combinatorial generalization on those patterns also worked well for these 3D objects. Thus, we are confident that the model's learning behavior is not idiosyncratic to the particular set of objects used here, and represents a general capacity of the system to develop abstract representations through predictive learning. Other ongoing work to be reported in an upcoming publication is applying the model to prediction of auditory speech inputs, which has a natural temporal structure, and finding similar results in terms of learning higher-level abstract encoding of these auditory signals.

To directly address the question of whether the hierarchical structure of the network supports the development of abstract, higher-level representations that go beyond the information present in the visual inputs, we applied a second-order similarity measure across the object-level similarity matrices computed at each layer in the network (Figure 6). This shows the extent to which the similarity matrix across objects in one layer is itself similar to the object similarity matrix in another layer, in terms of a correlation measure across these similarity matrices. Critically, this measure does not depend on any kind of subjective interpretation of the learned representations — it just tells us whether whatever similarity structure was learned differs across the layers. Starting from either V1 compared to all higher layers, or the highest TE layer compared to all lower layers, we found a consistent pattern of progressive emergence of the object categorization structure in the upper IT pathway (TEO, TE).

This analysis confirms that indeed the IT category structure is significantly different from that present at the level of the V1 primary visual input. Thus the model, despite being trained only to generate accurate visual input-level predictions, has learned to represent these objects in an abstract way that goes beyond the raw input-level information. We further verified that at the highest IT levels in the model, a consistent, spatially-invariant representation is present across different views of the same object (e.g., the average correlation across frames within an object was .901).

To better understand the nature of these learned representations, Figure 7 shows a representational similarity analysis (RSA) on the activity patterns at the highest IT layer (TE), which reveals the explicit categorical structure of the learned representations (Cadieu et al., 2014; Kriegeskorte, Mur, & Bandettini, 2008). Specifically, we found that the highest IT layer (TE) produced a systematic organization of the 156 3D objects into 5 categories. In our admittedly subjective judgment, these categories seemed to correspond to the overall shape of the objects, as shown by the object exemplars in the figure (pyramid-shaped, vertically-elongated, round, boxy / square, and horizontally-elongated). Furthermore, the basic-level categories were subsumed within these broader shape-level categories, so the model appears to be sensitive to the coherence of these basic-level categories as well, but apparently their shapes were not sufficiently distinct between categories to drive differentiated TE-level representations for each such basic-level category.

Given that the model only learns from a passive visual experience of the objects, it has no access to any of the richer interactive multi-modal information that people and animals would have. Furthermore, as evident in Figure 5b, the relatively low resolution of the V1 layers (required to make the model tractable computationally) means that complex visual details are not reliably encoded (and even so, are not generally reliable across object exemplars), such that the overall object shape is the most salient and sensible basis for categorization for this model.

Although these object shape categories appeared sensible to us, we ran a simple experiment to test whether a sample of 30 human participants would use the same category structure in evaluating the pairwise similarity of these objects. Figure 7b shows the results, confirming that indeed this same organization of the objects emerged in their similarity judgments. These judgments were based on the V1 reconstruction as shown in Figure 5b to capture the model's coarse-grained perception; see Appendix for methods and further analysis.

The progressive emergence of increasingly abstract category structure across visual areas, evident in Figure 6, has been investigated in recent comparisons between monkey electrophysiological recordings and deep convolutional neural networks (DCNNs), which provide a reasonably good fit the the overall progressive pattern of increasingly categorical organization (Cadieu et al., 2014). However, these DCNNs were trained on large datasets of human-labeled object categories, and it is perhaps not too surprising that the higher layers closer to these category output labels exhibited a greater degree of categorical organization. In contrast, because the only source of learning in our model comes from prediction errors over the V1 input layers, the graded emergence of an object hierarchy here reflects a truly self-organizing learning process.

Figure 8 compares the similarity structures in layers V4 and IT in macaque monkeys (Cadieu et al., 2014) with those in corresponding layers in our model. In both the monkeys and our model, the higher IT layer builds upon and clarifies the noisier structure that is emerging in the earlier V4 layer, showing that our model replicates the essential qualitative hierarchical progression in the brain. As noted, we would not expect our model to exactly replicate the detailed object-specific similarity structure found in macaques, due to the impoverished nature of our model's experience, so this comparison remains qualitative in terms of the respective differences between V4 and IT in each model, rather than a direct comparison of the similarity structure between corresponding layers in the model and the macaque. In the future, when we can scale up our model and tune the attentional processing dynamics necessary to deal with cluttered visual scenes, we will be able to train our model on the same images presented to the macaques, and can provide this more direct comparison.

Finally, we did not use analyses based on decoding techniques, because with high-dimensional distributed neural representations, it is generally possible to decode many different features that are not otherwise compactly and directly represented (Fusi, Miller, & Rigotti, 2016). In preliminary work using decoding in the context of the simpler feature-based input patterns, we indeed found that decoding was not a very sensitive measure of the differentiation of representations across layers, which is so clearly evident in Figure 6. Thus, as advocates of the RSA approach have argued, measuring similarity structure evident in the

activity patterns over a given layer generally provides a clearer picture of what that layer is explicitly encoding (Kriegeskorte et al., 2008).

In summary, the model learned an abstract category organization that reflects the overall visual shapes of the objects as judged by human participants, in a way that is invariant to the differences in motion, rotation, and scaling that are present in the V1 visual inputs. We are not aware of any other model that has accomplished this signature computation of the ventral *What* pathway in a purely self-organizing manner operating on realistic 3D visual objects, without any explicit supervised category labels. Furthermore, our model does this using a learning algorithm directly based on detailed properties of the underlying biological circuits in this pathway, providing a coherent overall account.

Backpropagation Comparison Models

To help discern some of the factors that contribute to the categorical learning in our model, and provide a comparison with more widely-used error backpropagation models, we tested a backpropagation-based (Bp) version of the same *What* vs. *Where* architecture as our biologically based predictive error model, and we also tested a standard *PredNet* model (Lotter et al., 2016) with extensive hyperparameter optimization (see Appendix). Due to the constraints of backpropagation, we had to eliminate any bidirectional connectivity loops in the Bp version, but we were able to retain a form of predictive learning by configuring the V1p pulvinar layer as the final target output layer, with the target being the next visual input relative to the current V1 inputs.

Figure 9 shows the same second-order similarity analysis as Figure 6, to determine the extent to which these comparison networks also developed more abstract representations in the higher layers that diverge from the similarity structure present in the lowest layers. According to this simple objective analysis, they did not — the higher layers showed no significant, progressive divergence in their similarity structure. The PredNet model did show a larger difference between the first layer and the rest of the layers, due to the subsequent layers encoding errors while the first layer has a positive representation of the image, but there was no progressive difference beyond that up into the higher layers.

Next, we examined the RSA matrices for the highest (TE) layer in the comparison models, also in comparison with the same for the V1 layer (Figure 10). This shows that the TE layer in the Bp model formed a simple binary category structure overall, which is similar to the RSA for the V1 input layer. It is also important to emphasize that the scales on these figures are different (as shown in their headers), such that these comparison models had much less differentiated representations overall. Similar results were found in the PredNet model. Because existing work with these models has typically relied on additional supervised learning and decoder-based analyses (which are essentially equivalent to an additional layer of supervised learning), these RSA-based analyses provide an important, more sensitive way of determining what they learn purely through predictive learning.

These results show that the additional biologically derived properties in our model are playing a critical role in the development of abstract categorical representations that go beyond the raw visual inputs. These properties include: excitatory bidirectional

connections, inhibitory competition, and an additional Hebbian form of learning that serves as a regularizer (similar to weight decay) on top of predictive error-driven learning (O'Reilly, 1998; O'Reilly & Munakata, 2000). Each of these properties could promote the formation of categorical representations. Bidirectional connections enable top-down signals to consistently shape lower-level representations, creating significant attractor dynamics that cause the entire network to settle into discrete categorical attractor states. Another indication of the importance of bidirectional connections is that a greedy layer-wise pretraining scheme, consistent with a putative developmental cascade of learning from the sensory periphery on up (Bengio, Yao, Alain, & Vincent, 2013; Hinton & Salakhutdinov, 2006; Shrager & Johnson, 1996; Valpola, 2014), did not work in our model. Instead, we found it essential that higher layers, with their ability to form more abstract, invariant representations, interact and shape learning in lower layers right from the beginning.

Furthermore, the recurrent connections within the TEO and TE layers likely play an important role by biasing the temporal dynamics toward longer persistence (Chaudhuri et al., 2015). By contrast, backpropagation networks typically lack these kinds of attractor dynamics, and this could contribute significantly to their relative lack of categorical learning. Hebbian learning drives the formation of representations that encode the principal components of activity correlations over time, which can help more categorical representations coalesce (and results below already indicate its importance). Inhibition, especially in combination with Hebbian learning, drives representations to specialize on more specific subsets of the space.

Ongoing work is attempting to determine which of these is essential in this case (perhaps all of them) by systematically introducing some of these properties into the backpropagation model, though this is difficult because full bidirectional recurrent activity propagation, which is essential for conveying error signals top-down in the biological network, is incompatible with the standard efficient form of error backpropagation, and requires significantly more computationally intensive and unstable forms of fully recurrent backpropagation (Pineda, 1987; Williams & Zipser, 1992). Furthermore, Hebbian learning requires dynamic inhibitory competition which is difficult to incorporate within the backpropagation framework.

Architecture and Parameter Manipulations

Figure 11 shows just a few of the large number of parameter manipulations that have been conducted to develop and test the final architecture. For example, we hypothesized that separating the overall prediction problem between a spatial *Where* vs. non-spatial *What* pathway (Goodale & Milner, 1992; Ungerleider & Mishkin, 1982), would strongly benefit the formation of more abstract, categorical object representations in the *What* pathway. Specifically, the *Where* pathway can learn relatively quickly to predict the overall spatial trajectory of the object (and anticipate the effects of saccades), and thus effectively regress out that component of the overall prediction error, leaving the residual error concentrated in object feature information, which can train the ventral *What* pathway to develop abstract visual categories.

Figure 11a shows that, indeed, when the *Where* pathway is lesioned, the formation of abstract categorical representations in the intact *What* pathway is significantly impaired. We also hypothesized that full predictive learning (about the future), as compared to just encoding and decoding the current state (i.e., an auto-encoder, which is much easier computationally), is also critical for the formation of abstract categorical representations — prediction is a "desirable difficulty" (Bjork, 1994). Figure 11b shows that this was the case. Finally, consistent with our hypothesis that Hebbian learning provides an important bias on learning, Figure 11c shows the impairment associated with reducing this learning bias. The significant reduction in differentiation across all of these manipulations shows that this differentiation property is not a simple consequence of the neural architecture, but rather depends critically on the learning process, unfolding over time with appropriate parameter values and other architectural components. Furthermore, the Bp comparison model shares the same architecture, and does not show the differentiation across layers.

Predictive Behavior

A signature example of predictive behavior at the neural level in the brain is the *predictive remapping* of visual space in anticipation of a saccadic eye movements (Colby, Duhamel, & Goldberg, 1997; Duhamel et al., 1992; Gottlieb, Kusunoki, & Goldberg, 1998; Marino & Mazer, 2016; Nakamura & Colby, 2002) (Figure 12a). Here, parietal neurons start to fire at the *future* receptive field location where a currently-visible stimulus will appear after a planned saccade is actually executed. Remapping has also been shown for border ownership neurons in V2 (O'Herron & von der Heydt, 2013) and in area V4 (Neupane, Guitton, & Pack, 2016, 2020). These are examples, we believe, of a predictive process operating throughout the neocortex to predict what will be experienced next. A major consequence of this predictive process is the perception of a stable, coherent visual world despite constant saccades and other sources of visual change.

Figure 12b shows that our model exhibits this predictive remapping phenomenon. Specifically, LIP, which is most directly interconnected with the saccade efferent copy signals, is the first to predict the new location, and it then drives top-down activation of lower layers. This top-down dynamic is consistent with the account of predictive remapping given by Wurtz (2008) and Cavanagh et al. (2010), who argue that the key remapping takes place at the high levels of the dorsal stream, which then drive top-down activation of the predicted location in lower areas, instead of the alternative where lower-levels remap themselves based on saccade-related signals. The lower-level visual layers are simply too large and distributed to be able to remap across the relevant degrees of visual angle — the extensive lateral connectivity needed to communicate across these areas would be prohibitive.

Neural Data and Predictions

Having tested the computational and functional learning properties of this biologically based predictive learning mechanism, we now return to consider some of the most important neural data of relevance to our hypotheses, beyond that summarized in the introduction, including contrasts with a widely-discussed alternative framework for predictive coding,

and some of the extensive data on alpha frequency effects, followed by a discussion of predictions that would clearly test the validity of this framework.

Additional Neuroscience Data

We begin with data relevant to the basic neural-level properties of the framework. First, a central element of the proposed model is the alpha cycle bursting, and subsequent inter-burst pauses, in the 5IB neurons. Direct electrophysiological recording of deep layer neurons shows periodic alpha-scale bursting for continuous tones *in awake animals* (Luczak et al., 2009, 2013; Sakata & Harris, 2009, 2012). *In vitro*, a variety of potential mechanisms behind the generation and synchronization of the 5IB bursts driving this alpha cycle have been identified (Connors et al., 1982; Franceschetti et al., 1995; Silva et al., 1991). Furthermore, the pulvinar has been shown to drive alpha-frequency synchronization of cortical activity across areas in the alpha band in awake behaving animals (Saalmann et al., 2012). We review the larger alpha frequency literature in more detail below, but it is critical to emphasize that this alpha bursting dynamic is actually found in awake, behaving animals, because so many other bursting and up / down state phenomena have recently been shown to only occur in anesthetized brains, including bursting in the thalamic TRC neurons.

In contrast to the 5IB bursting, the 6CT neurons exhibit regular spiking behavior, (Thomson, 2010; Thomson & Lamy, 2007), providing consistent activation to the pulvinar. Also, they do not have axonal branches that project to other cortical areas — the subpopulation that projects to the pulvinar only project there and not to other cortical areas (Petrof, Viaene, & Sherman, 2012), whereas there are other layer 6 neurons that do project to other cortical areas. This distinct connectivity is consistent with a specific role of this neuron type in generating predictions in the pulvinar. The 6CT synaptic inputs on pulvinar TRCs have metabatropic glutamate receptors (mGluR) that have longer time-scale temporal dynamics consistent with the alpha period (100 ms) and even longer (S. M. Sherman, 2014), and the 6CT neurons themselves also have temporally-delayed responding (Harris & Shepherd, 2015; Sakata & Harris, 2009; Thomson, 2010). Furthermore, they have significantly more plasticity-inducing NMDA receptors compared to the 5IB projections (Usrey & Sherman, 2018). These properties are consistent with the 6CT inputs driving a longer-integrated prediction signal that is subject to learning, whereas the 5IB are likely non-plastic and their effects are tightly localized in time.

The 5IB inputs often have distinctive *glomeruli* structures at their synapses onto pulvinar neurons, which contain a complete feedforward inhibition circuit involving a local inhibitory interneuron, in addition to the direct strong excitatory driver input (Wilson, Bose, Sherman, & Guillery, 1984). Computationally, this can provide a balanced level of excitatory and inhibitory drive so as to not overly excite the receiving neuron, while still dominating its firing behavior.

Although there are well-documented and widely-discussed burst vs. tonic firing modes in pulvinar neurons (S. M. Sherman & Guillery, 2006), there is not much evidence of these playing a clear role in the awake, behaving state, and as noted earlier the growing electrophysiological evidence shows a remarkable correspondence between cortical and pulvinar response properties across multiple different pulvinar areas in this awake state.

Nevertheless, there may be important dynamics arising from these firing modes that are more subtle or emerge in particular types of state transitions that may have yet to be identified.

Contrast with Explict Error (EE) Frameworks

To further clarify the nature of the present theory, and introduce a body of relevant data, we contrast it with the widely-discussed explicit error (*EE*) framework for predictive coding (Bastos et al., 2012; Friston, 2005, 2010; Kawato et al., 1993; Lotter et al., 2016; Ouden et al., 2012; Rao & Ballard, 1999) (Figure 13). The hypothesized locus for computing errors in this framework is in the superficial layers of the neocortex, which are suggested to directly compute the difference between bottom-up inputs from lower layers and top-down inputs from higher areas. Despite many attempts to identify such explicit error-coding neurons in the cortex, no substantial body of unambiguous evidence has been discovered (Kok & de Lange, 2015; Kok, Jehee, & de Lange, 2012; Lee & Mumford, 2003; Summerfield & Egner, 2009; Walsh, McGovern, Clark, & O'Connell, 2020). Furthermore, due to the positive-only firing rate nature of neural coding, two separate populations would be required to convey both signs of prediction error signals, or it would have to be encoded as a variation from tonic firing levels, which are generally low in the neocortex.

By contrast, the use of temporal-difference error signals enables all connections between cortical layers to be excitatory and each layer can represent the positive encoding of either the prediction or outcome state, at different levels of abstraction. These properties are overwhelmingly supported by extensive electrophysiological data about the hierarchical organization of representations, e.g., in the visual object recognition pathway (Cadieu et al., 2014; Kobatake & Tanaka, 1994; VanRullen & Thorpe, 2002), and are consistent with the widely-supported biased competition model for excitatory top-down attentional effects (Desimone & Duncan, 1995; E. K. Miller & Cohen, 2001; O'Reilly et al., 2013; Reynolds et al., 1999).

The EE approach requires net inhibitory top-down predictions, and it sends error signals forward, not positive representations of the actual state at a given level of abstraction. Thus a literal interpretation (and at least one existing implementation; Lotter et al., 2016) has only error signals represented at all levels above the lowest level, which is inconsistent with the positive encoding of stimuli at various levels of abstraction across the visual hierarchy. For example, although Issa, Cadieu, and DiCarlo (2018) observed an error-signal-like increase in activation for atypical faces in some pIT neurons, these neurons overall had a positive stimulus encoding, with only a relatively small, later, error-like modulation.

Furthermore, as discussed below, anticipatory predictions typically closely resemble the subsequent stimulus-driven activity, suggesting a positive, not inhibitory, effect (Cavanagh et al., 2010; Duhamel et al., 1992; Lee & Mumford, 2003; Walsh et al., 2020). However, there are various different ways of reformulating the neural implementation of EE that can avoid some of these issues (Bastos et al., 2012; Spratling, 2008), but perhaps this flexibility renders the framework difficult to falsify (Kogo & Trengove, 2015). In any case, an extensive treatment of the issues with EE is beyond the scope of this paper and has already been aptly covered by Walsh et al. (2020) — our goal here is to highlight some of

the core differences as a way to clarify the framework by way of contrast, and in relation to available data.

First, there are many examples of anticipatory predictive neural firing in the brain. Of perhaps greatest relevance, Barczak et al. (2018) recently showed that the auditory pulvinar in monkeys exhibits predictive firing using a carefully controlled auditory sequence that had no first-order acoustic differences from a background noise signal. The pulvinar predictive activation preceded that of A1, suggesting a strong predictive role for pulvinar. Unfortunately, the deep layers of higher auditory areas that should contribute to the formation of the pulvinar prediction were not recorded in this study, so their role in generating the prediction could not be determined.

Nevertheless, there is extensive additional evidence for top-down anticipatory activation of predicted stimuli, with activity patterns closely resembling the subsequent stimulus-driven ones (Walsh et al., 2020). For example, the widely replicated predictive remapping effect, simulated in our model (Figure 12) is of this nature (Cavanagh et al., 2010; Duhamel et al., 1992; Wurtz, 2008). The fact that these anticipatory activations are of a positive nature, consistent with the stimulus-driven activations, is inconsistent with the expected behavior of EE neurons, which should be inhibited by the top-down prediction, while not receiving any bottom-up stimulus.

However, the neural response to the actual predicted stimulus itself is typically suppressed relative to unexpected stimuli, i.e., *expectation suppression* (Bastos et al., 2012; Meyer & Olson, 2011; Summerfield et al., 2008; Todorovic et al., 2011). This phenomenon is widely cited as evidence in favor of the EE predictive coding framework, consistent with an inhibitory effect of the expectation. Nevertheless, despite various conflicting results and many complications of interpretation, multiple comprehensive reviews conclude that it is difficult to distinguish expectation suppression from the neural adaptation effects that underlie the well-documented *repetition suppression* effect (Kok & de Lange, 2015; Kok et al., 2012; Lee & Mumford, 2003; Summerfield & Egner, 2009; Vinken & Vogels, 2017; Walsh et al., 2020). Furthermore, detailed single-neuron level recordings are the least likely to show these effects — instead, they are most evident in aggregate signals such as the BOLD response in fMRI, suggesting that they may more strongly reflect population-level differences in activity, rather than individual explicit error coding neurons.

As noted earlier, accurately predicted outcomes in our framework would result in a continued adaptation of the neural response carrying over from the prediction to the outcome state, whereas unexpected outcomes would be associated with two distinct patterns of activity over a given area: first the prediction and then the outcome. Thus, the unexpected outcome state would not be subject to the prior neural adaptation effects, and furthermore the time-integrated aggregate activity over these two patterns would be greater compared to the single activity state associated with an accurately predicted outcome. Thus, our model explains expectation suppression without invoking EE neurons, meaning that considerably more detailed and replicable experimental paradigms using single-neuron resolution techniques are needed to distinguish EE from our framework.

Alpha Frequency Effects

The alpha frequency bursting of 5IB neurons acting as drivers into the pulvinar naturally entrains the predictive learning process in our model to this fundamental rhythm, which has long been recognized as an important signature of posterior cortical function (Berger, 1929; Nunn & Osselton, 1974; VanRullen & Koch, 2003; Varela, Toro, John, & Schwartz, 1981; Walter, 1953). A number of different functional associations with alpha have been established, and this literature is large and growing rapidly. Thus, we refer the reader to recent reviews (Clayton et al., 2018; Foster & Awh, 2019; Jensen, Bonnefond, Marshall, & Tiesinga, 2015; VanRullen, 2016) while highlighting the data most relevant to our specific framework here, organized according to a set of key points.

- *Alpha is specifically associated with deep neocortical layers and the pulvinar, and with feedback pathways in the cortex.* This has been established using direct laminar-specific electrophysiological single-neuron and local field potential (LFP) recordings (Buffalo et al., 2011; Luczak et al., 2013; Maier, Adams, Aura, & Leopold, 2010; Maier, Aura, & Leopold, 2011; Spaak, Bonnefond, Maier, Leopold, & Jensen, 2012; Xing, Yeh, Burns, & Shapley, 2012), and feedforward vs. feedback manipulations (Bastos et al., 2015; Jensen et al., 2015; Michalareas et al., 2016; van Kerkoerle et al., 2014; von Stein, Chiang, & König, 2000). These data are consistent with the 5IB alpha bursting and thë major role of cortical deep layers in driving top-down corticocortical projections (in addition to the 6CT pathway which is specific to the pulvinar). By contrast, these same papers show that superficial cortical layers are associated with gamma frequency (40 Hz) dynamics. However, the next point raises some important interpretational difficulties.
- *Increases in cortical activity levels, e.g., due to attention, produce a corresponding decrease in alpha power, while decreased activity increases alpha power* (Foster & Awh, 2019; Fries, Womelsdorf, Oostenveld, & Desimone, 2008; Jensen & Mazaheri, 2010; Kelly, Lalor, Reilly, & Foxe, 2006; Klimesch, Sauseng, & Hanslmayr, 2007; Worden, Foxe, Wang, & Simpson, 2000). This pattern is not exactly what you might expect if alpha was a signature of predictive learning. However, given that these same pulvinar and thalamocortical pathways are also widely regarded as important for attention (Bender & Youakim, 2001; Fiebelkorn & Kastner, 2019; LaBerge & Buchsbaum, 1990; Saalmann & Kastner, 2011; Snow et al., 2009; Zhou et al., 2016), this pattern presents a challenge for many theorists. However, it is possible to explain this pattern as arising directly from the desynchronizing effects of cortical activity on alpha power. Specifically, neural spiking is associated with broadband noise, due to the highly random, Poisson nature of spike firing, which can desynchronize the entrainment of lower-frequency oscillations including alpha (Privman, Malach, & Yeshurun, 2013; Ray & Maunsell, 2011; Solomon et al., 2017; Waldert, Lemon, & Kraskov, 2013). In other words, because cortical activity is inherently noisy, it tends to interfere with the coherent activity across populations of neurons needed to produce a strong alpha frequency power signal. This explanation is directly supported by studies manipulating and measuring

cortical activity (Fries et al., 2008; Zhou et al., 2016), and is consistent with alpha power changes being a *result* of attentional modulation, but not their cause (Antonov, Chakravarthi, & Andersen, 2020). Thus, while attention and predictive learning can both affect overall activity levels in cortex, and thus drive changes in alpha power, alpha power itself is not a transparent measure of the underlying mechanisms supporting these functions, which may help to explain some contradictory patterns of results (Foster & Awh, 2019; Gundlach, Moratti, Forschack, & Müller, 2020; Keitel et al., 2019).

- *Alpha phase effects provide a more direct measure of thalamocortical function than alpha power, and have been more consistently related to perception, attention, and prediction* (Busch, Dubois, & VanRullen, 2009; Jaegle & Ro, 2013; K. E. Mathewson, Fabiani, Gratton, Beck, & Lleras, 2010; Neupane et al., 2017; Nunn & Osselton, 1974; Palva & Palva, 2011; Solís-Vivanco, Jensen, & Bonnefond, 2018; VanRullen & Koch, 2003; Varela et al., 1981). For example, weak, near-threshold stimuli are more reliably detected and processed when presented in the trough of the individual's ongoing alpha cycle. Of greatest relevance to the present paper are studies showing effects of prediction on alpha phase (Mayer, Schwiedrzik, Wibral, Singer, & Melloni, 2016; Samaha, Bauer, Cimaroli, & Postle, 2015; M. T. Sherman, Kanai, Seth, & VanRullen, 2016). For example, Mayer et al. (2016) showed that prestimulus alpha phase directly correlated with the predictability of the upcoming stimulus, and the pattern of this prestimulus activation was indistinguishable from the subsequent stimulus activation pattern. This is consistent with our model, and less consistent with the EE framework, as discussed previously. Neupane et al. (2017) found strong alpha coherence effects in LFP recordings distributed across V4, associated with the predictive remapping of receptive fields (Duhamel et al., 1992).
- *Discrete, salient, or oscillatory stimuli entrain the alpha cycle in the brain* (K. E. Mathewson et al., 2012; Spaak, de Lange, & Jensen, 2014). Furthermore, the massive literature on *event related potentials* (ERPs) may represent a significant contribution from alpha-level entrainment (Gruber, Klimesch, Sauseng, & Doppelmayr, 2005; Klimesch, 2011; Makeig et al., 2002). These entrainment effects are consistent with the 5IB entrainment mechanisms in our framework, as described earlier, and entrainment is functionally important for aligning predictive learning with relevant salient or unexpected outcomes.
- *The pulvinar contributes to synchronizing alpha phase relationships across different brain areas* (Fiebelkorn, Pinsk, & Kastner, 2018; Saalmann et al., 2012). This is consistent with the broad, convergent pattern of projections into the pulvinar from many different cortical areas, and the corresponding broad projections back out to these same areas (Arcaro et al., 2015; Shipp, 2003). Functionally, this convergence and synchronization is important for integrating the contributions from these different areas at the same time, to generate predictions over the pulvinar.

- *The theta cycle, comprised of a pair of alpha cycles, organizes saccades, and attentional, motor, and mnemonic processes* (Fiebelkorn & Kastner, 2019). The theta rhythm is dominant in the medial temporal lobe and hippocampus, and has been extensively studied there (Buzsáki, 2005; Kahana, Seelig, & Madsen, 2001). Furthermore, there is a sharp peak of saccade fixation durations at 200 ms, which suggests that two alpha cycles are typically required for complete processing of a given fixation. On the first cycle, the predictions from before the eye moved may be fairly vague depending on factors such as the size of the saccade and familiarity with the environment. But after the first alpha cycle of a fixation, a subsequent *postdiction* phase can provide an important additional learning opportunity, to consolidate and more deeply encode the current fixation (computationally equivalent to an auto-encoder). Also, a mix of smaller saccades (including microsaccades) and larger saccades enables a range of more and less predictable outcomes on the first alpha cycle after the saccade, and matches human behavior (Martinez-Conde, Macknik, & Hubel, 2004; Martinez-Conde, Otero-Millan, & Macknik, 2013).

Putting all of these points together, a particularly effective way of testing the predictions of our framework would be measuring alpha phase changes emerging in the prestimulus period as a function of predictive learning in predictable sequential stimulus streams. In addition, it would also be important to examine theta and alpha-cycle dynamics in relation to predictive learning in the context of attention, motor control, and memory processes, to better understand the larger systems-level temporal organization of learning and processing in the brain (Fiebelkorn & Kastner, 2019).

Predictions for Predictive Learning

In this section, we enumerate a set of direct, testable predictions from our framework. Before doing so, there are several important considerations for any experimental test of the theory. First, the nature of what is to be learned must be matched to the pulvinar area in question. For example, learning a new variation of basic physics in movies at the alpha time scale (e.g., altering properties such as gravity, inertia, or elasticity), would be appropriate for the lower level visual pathways. At higher visual levels (e.g., IT cortex), it might be possible to use simple sequences of different objects, although it is not clear to what extent the hippocampus or prefrontal cortex might also contribute in this case (Fiser et al., 2016; Gavornik & Bear, 2014). To distinguish pulvinar learning effects from pervasive motor learning supported by other brain areas, it would be most effective to directly measure activity in the pulvinar and / or associated perceptual neocortical areas, instead of involving overt behavioral performance.

Much of the learning in posterior sensory cortex should take place early in development, requiring very early developmental interventions or genetic knockouts that are expressed from the start (which can also have other interpretational issues if not highly selective). In our models, the bulk of the basic sensory predictive learning happens very quickly, because the basic first-level regularities are quite strong and relatively easily learned. While there are longer-term changes in the higher-level pathways in our models, more fine-grained measurements would likely be required to see these changes. Once this learning has taken

place, the remaining contributions of the thalamocortical circuit are likely more strongly weighted toward its role in attention, as we discuss below. Finally, directly lesioning or inactivating the pulvinar is not likely to be very informative, because existing work has shown dramatic effects on cortical activity (Purushothaman, Marion, Li, & Casagrande, 2012; Zhou et al., 2016), and also any effects could be attributed to the attentional contributions of the pulvinar.

With these considerations in mind, here are a set of strong predictions from our model that should be testable using existing techniques. Failure to obtain the predicted result, while adhering to all the relevant constraints, would constitute a falsification of our model.

- *Blocking 5IB bursting mechanisms early in developmental learning should disrupt learning*. It should be possible to selectively knock out or modify the channels that cause this specific population of neurons to burst fire, and doing so should have a significant effect on learning in associated neocortical and pulvinar areas, given the critical role that this burst firing plays on the predictive learning process as elaborated above.
- *Blocking synaptic plasticity in pulvinar (specifically the 6CT inputs) very early in developmental learning should impair learning*. While most of the learning overall should occur in the neocortex as a result of the temporal difference error signal broadcast by the pulvinar (which should remain generally intact), learning in the 6CT projections is important, especially right at the start, to map the emerging neocortical representations into the space defined by the 5IB projections.
- *Temporal differences on an alpha cycle timescale actually drive synaptic plasticity in an error-driven learning manner, in neocortical pyramidal neurons and in 6CT inputs to pulvinar*. That is, if a pre / post pair of neurons across a synapse is more active in the prediction than the subsequent outcome, the synapse should experience LTD (long term depression), and vice-versa if the activity pattern is reversed (long term potentiation, LTP, for more activity in outcome than prediction). Furthermore, if activity is essentially stable across both prediction and outcome phases, then weights should not change (modulo a small level of Hebbian learning; O'Reilly & Munakata, 2000; O'Reilly et al., 2012). This should be directly testable using current experimental methods, and is perhaps the single most important empirical test of this entire framework, and it also underlies many other current approaches to error-driven learning in the brain (Bengio et al., 2017; Lillicrap et al., 2020; Whittington & Bogacz, 2019). One general consideration is the extent to which an awake *in vivo* preparation would be required to capture all the neuromodulatory and other factors present when this learning normally takes place. Some suggestive evidence in such a preparation is generally consistent with a sensitivity to relatively short-term temporal dynamics (Lim et al., 2015), although these results lacked the direct measurement of individual neural activity across a synapse.

Discussion

We have hypothesized a novel computational function for the distinctive features of thalamocortical circuits (S. M. Sherman & Guillery, 2006; Usrey & Sherman, 2018), as supporting a specific form of prediction-error driven learning, where predictions arise from the numerous top-down layer 6CT projections into the pulvinar, and the strong, sparse driving 5IB inputs supply the bottom-up sensory-driven outcome. The phasic bursting nature of the 5IB inputs results in a natural temporal-difference error signal of prediction followed by outcome, consistent with extensive neural recording data. This temporal dynamic is also essential for enabling predictions to be generated without contamination from current sensory inputs, and predicts a characteristic alpha frequency prediction cycle based on the 10hz bursting cycle of the 5IB inputs, consistent with the pervasive influence of alpha on perception and neural dynamics (Clayton et al., 2018; Foster & Awh, 2019; Jensen et al., 2015; VanRullen, 2016). In short, the hypothesized predictive learning function fits remarkably well with a number of well-established properties of these thalamocortical circuits, and we also provided a set of additional predictions that could be tested to further evaluate this theory, especially in contrast to the widely-discussed alternative of explicit error coding neurons, which have not been unambiguously supported across a range of empirical studies (Walsh et al., 2020).

Furthermore, we implemented this theory in a large scale model of the visual system, and demonstrated that learning based strictly on predicting what will be seen next is, in conjunction with a number of critical biologically motivated network properties and mechanisms, capable of generating abstract, invariant categorical representations of the overall shapes of objects. The nature of these shape representations closely matches human shape similarity judgments on the same objects. Thus, predictive learning has the potential to go beyond the surface structure of its inputs, and develop systematic, abstract encodings of the environment. We found that comparison models based on standard error backpropagation learning did not learn a categorical structure that went beyond the surface similarity present in the visual input layers, and future work is focused on narrowing down the specific mechanisms required to drive this learning.

In addition to the predictive learning functions of the deep / thalamic layers, these same circuits are also likely critical for supporting powerful top-down attentional mechanisms that have a net multiplicative effect on superficial-layer activations (Bortone, Olsen, & Scanziani, 2014; Bortone et al., 2014; Olsen, Bortone, Adesnik, & Scanziani, 2012; Olsen et al., 2012). The importance of the pulvinar for attentional processing has been widely documented (Bender & Youakim, 2001; LaBerge & Buchsbaum, 1990; Saalmann et al., 2012, e.g.,), and there is likely an additional important role of the thalamic reticular nucleus (TRN), which can contribute a surround-inhibition contrast-enhancing effect on top of the incoming attentional signal from the cortex (Crick, 1984; Jaramillo, Mejias, & Wang, 2019; Pinault, 2004; Wimmer et al., 2015). In other work in progress, we have shown that the deep / thalamic circuits in our model produce attentional effects consistent with the abstract Reynolds and Heeger (2009) model, while the contributions of the deep layer networks to this function are broadly consistent with the folded-feedback model (Grossberg, 1999). These attentional modulation signals cause the bidirectional constraint satisfaction process

in the superficial network to focus on task-relevant information while down-regulating responses to irrelevant information — in the real world, there are typically too many objects to track at any given time, so predictive learning must be directed toward the most important objects (Cavanagh et al., 2010; Pylyshyn, 1989; Richter & de Lange, 2019).

There is also data suggesting that the pulvinar is important for supporting *confidence* judgments, driven by relative ambiguity in a random dot motion categorization task (Komura et al., 2013). Critically for the present framework, this confidence modulation only emerged in the period after the first 100 ms of processing, and manifested as a positive correlation with confidence (i.e., more unambiguous stimuli resulted in higher firing rates). We can interpret this as reflecting an ongoing generative *postdiction* of the stimulus signal, with stronger firing associated with more unambiguous top-down activation based on the current internal representation. Note that this directionality is the opposite of explicit error-coding neurons, which would presumably increase with increasing error / ambiguity in the prediction. Interestingly, inactivation of these pulvinar neurons resulted in a substantial (200%) increase in opt-out choices on the most ambiguous stimuli, suggesting a level of metacognitive awareness of the pulvinar signal (or at least a direct effect of pulvinar on relevant metacognitive processes). Predictive accuracy would be an ideal source of metacognitive confidence signals across a wide range of domains, suggesting another important contribution of pulvinar even after initial learning. Jaramillo et al. (2019) present a comprehensive model of attentional, decision-making, and working memory contributions of the pulvinar, including this confidence data, which is generally compatible with our framework, although it does not address any learning phenomena.

There are a number of important limitations of the current WWI model, in terms of its scale and ability to process real-world cluttered visual scenes with multiple objects present, such as those used in the widely-studied ImageNet dataset. The model is much smaller than standard DCNN vision models, because its computational demands are significantly higher, in a way that also does not fit well with current GPU-based parallel computation hardware, due to the relative complexity of the algorithms and the sparseness of the activations. For each image, 100 cycles (of 1 ms each) of activation updating are required to enable the bidirectional activation and inhibition to integrate in a graded manner over the alpha cycle, compared to just 1 such iteration for most feedforward DCNN models. Furthermore, the bidirectional connectivity, extensive shortcut connections, and use of multiple cortical lamina per cortical area result in significant increases in the number of synaptic connections, which dominate the computational cost, and scale roughly as n^2 in the number of neurons n per layer across one projection. Thus, there are 207 million connections for the full WWI model, requiring 10 GB of RAM, and it takes over a day to run using 32 high-performance CPU processors with fast network interconnects, using the fastest combination of threading and parallel batch training. Doubling the network size causes it to no longer fit in available RAM, and yet its high-resolution V1 layer is only 16×16, compared to 55×55 for basic DCNN models such as AlexNet, and 224×224 for VGG16. The result is that the model has a relatively low resolution view of the world, as reflected in the reconstructed images shown in Figure 5.

In addition to having a higher-resolution input to be able to process more complex real-world cluttered images, the model would require functional attentional dynamics to focus processing on a small number of objects at a time, as is well-documented for humans processing complex images. Thus, once the attentional dynamics are well-integrated with the predictive learning mechanisms, we can begin to explore performance on more complex images, subject to improved computational hardware supporting larger network sizes.

Considerable further work remains to be done to more precisely characterize the essential properties of our biologically motivated model necessary to produce this abstract form of learning, and to further explore the full scope of predictive learning across different domains. We strongly suspect that extensive cross-modal predictive learning in real-world environments, including between sensory and motor systems, is a significant factor in infant development and could greatly multiply the opportunities for the formation of higher-order abstract representations that more compactly and systematically capture the structure of the world (Yu & Smith, 2012). Future versions of these models could thus potentially provide novel insights into the fundamental question of how deep an understanding a pre-verbal human, or a non-verbal primate, can develop (J. Elman et al., 1996; Spelke, Breinlinger, Macomber, & Jacobson, 1992), based on predictive learning mechanisms. This would then represent the foundation upon which language and cultural learning builds, to shape the full extent of human intelligence.

Acknowledgments

We thank Dean Wyatte, Tom Hazy, Seth Herd, Kai Krueger, Tim Curran, David Sheinberg, Lew Harvey, Jessica Mollick, Will Chapman, Helene Devillez, and the rest of the CCN Lab for many helpful comments and suggestions. Supported by: ONR grants ONR N00014-19-1-2684 / N00014-18-1-2116, N00014-14-1-0670 / N00014-16-1-2128, N00014-18-C-2067, N00014-13-1-0067, D00014-12-C-0638. This work utilized the Janus supercomputer, which is supported by the National Science Foundation (award number CNS-0821794) and the University of Colorado Boulder. The Janus supercomputer is a joint effort of the University of Colorado Boulder, the University of Colorado Denver and the National Center for Atmospheric Research. All data and materials will be available at https://github.com/ccnlab/deep-obj-cat upon publication.

Appendix

All of the materials described here, including the experimental study, the computational models, and the code to perform the representational similarity analysis, are all available on our github account at: https://github.com/ccnlab/deep-obj-cat and the new version of the *emergent* simulation environment is at: https://github.com/emer/leabra which contains extensive documentation and examples that can be run in Python or the Go language. The best place to start in understanding computationally how the predictive learning model works is with the FSA model described in the main text, which is available at: https://github.com/emer/leabra/tree/master/examples/deep_fsa. For the large and complex WWI model, the most complete understanding can only be had by directly examining the code, as there are a number of details that are not efficiently captured in this Appendix text.

Representational Similarity Analysis Methods

The different representations being compared here are:

Leabra: The DeepLeabra (biological model) TE layer representations (specifically TEs = superficial – results are very similar for deep as well).

Bp: The TEs layer representations from the backpropagation version of biological model, including *What, Where* and *What * Where* integration layers, trained with the V1p and V1hp (low and high resolution pulvinar) layers as final output layers, using the time t target pattern from the $t - 1$ input (i.e., as a predictive network).

V1: The gabor-filtered representation of the visual input to both of the above models, which was identical across them.

PredNet: Highest layer (6th Layer) of the PredNet architecture.

Expt: Similarity matrix constructed from human pairwise similarity judgments (see Behavioral Experiment Methods).

An optimal category cluster can be defined as one that has high within-cluster similarity and low between-cluster similarity. This can be operationalized by the *contrast* distance metric, based on a 1-correlation (*dissimilarity*) measure, as the difference between the average within-cluster similarity and the average between-cluster similarity:

$$cd = \langle 1 - r_{in} \rangle - \langle 1 - r_{out} \rangle \quad (1)$$

With distance-like 1-correlation values, this contrast distance should be minimized (it is typically negative), or equivalently the contrast on raw correlation values can be maximized (it is typically a positive number – just the sign flip of distance value). We refer to the positive numbers and maximization here as that is more intuitive.

Starting with an initial set of clusters, a permutation-based hill-climbing strategy was used to determine a local minimum in this measure: each item was tested in each of the other possible categories, and if that configuration reduced the overall average contrast distance metric across all items, then it was adopted and the process iterated until no such permutation improved the metric. This algorithm can only decrease the number of clusters (by moving all items out of a given cluster), so different numbers of initial clusters can be used to search the overall space.

Figure 14 shows the resulting categories. The Bp model converged on the same cluster state from all starting configurations tested, varying from 5 to 2 initial categories. This is the cluster set shown in Figure 10 of the main paper, and has an average contrast distance (*acd*) of 0.0838 (this is relatively low because the patterns were overall quite similar). Likewise, the V1 patterns (which were the same across Leabra and Bp models) reliably converged on the same pattern (shown in Figure 10), with *acd* = 0.2448.

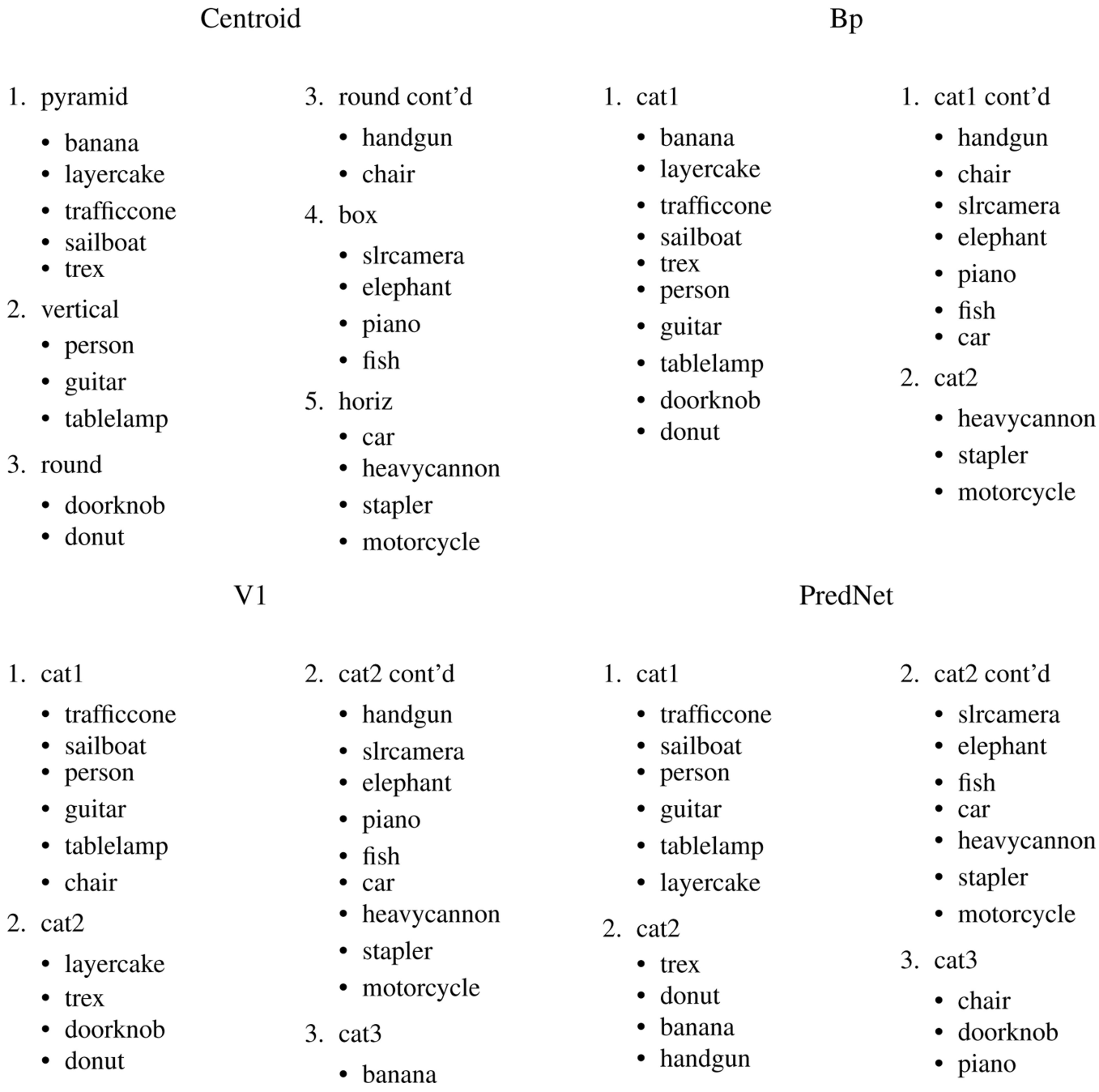

Figure 14:

Shape categories used for similarity matrix plots in main paper. *Centroid* shape categories are near-best for both the Leabra model and the Expt results, and fit our visual intuitions about overall shape. *Bp* are reliably optimal for Bp model from all starting points. *V1* are reliably optimal for V1 inputs, and also were close to the best for the Bp and PredNet layer 6 representations. *PredNet* are best stable solution for PredNet layer 6.

For the PredNet layer 6 representations, starting from the V1 categories gave the best results of any other set (*acd* = 0.1967), and a few permutations resulted in a reliable solution that

was arrived at from all other 3 category starting points tested, shown in Figure 14 (*acd* = 0.2820). This indicates that PredNet did not go much beyond the structure present in the input, even though it did not use the V1 gabor filtering used in the Leabra and Bp models (i.e., this V1-level encoding well-captures the structure of the visual inputs in general). The PredNet pixel and layer 1 representations both converged on essentially a single monolithic category with very low acd (0.0018, 0.0013).

For the Leabra TE representations, we found a set of *centroid* shape categories that are near-best when considering both the Leabra model and the results from the human behavioral experiment (Expt). Starting from these categories, the permutation analysis converged on reducing the size of the vertical and round categories to one item each, over a sequence of 5 steps. This is consistent with the observation from Figure 7 that there are three broader categories within which the 5 finer-grained categories are embedded (i.e., vertical and pyramid are overall similar to each other, as are round and box). Nevertheless, our initial visual intuition about the broad shape categories, along with a bias against having single-item categories, reinforced the use of the finer-grained centroid selection. The average contrast difference of our centroid selection is 0.5071, while the maximal result from the permutation was 0.5526, which is a relatively small proportional difference.

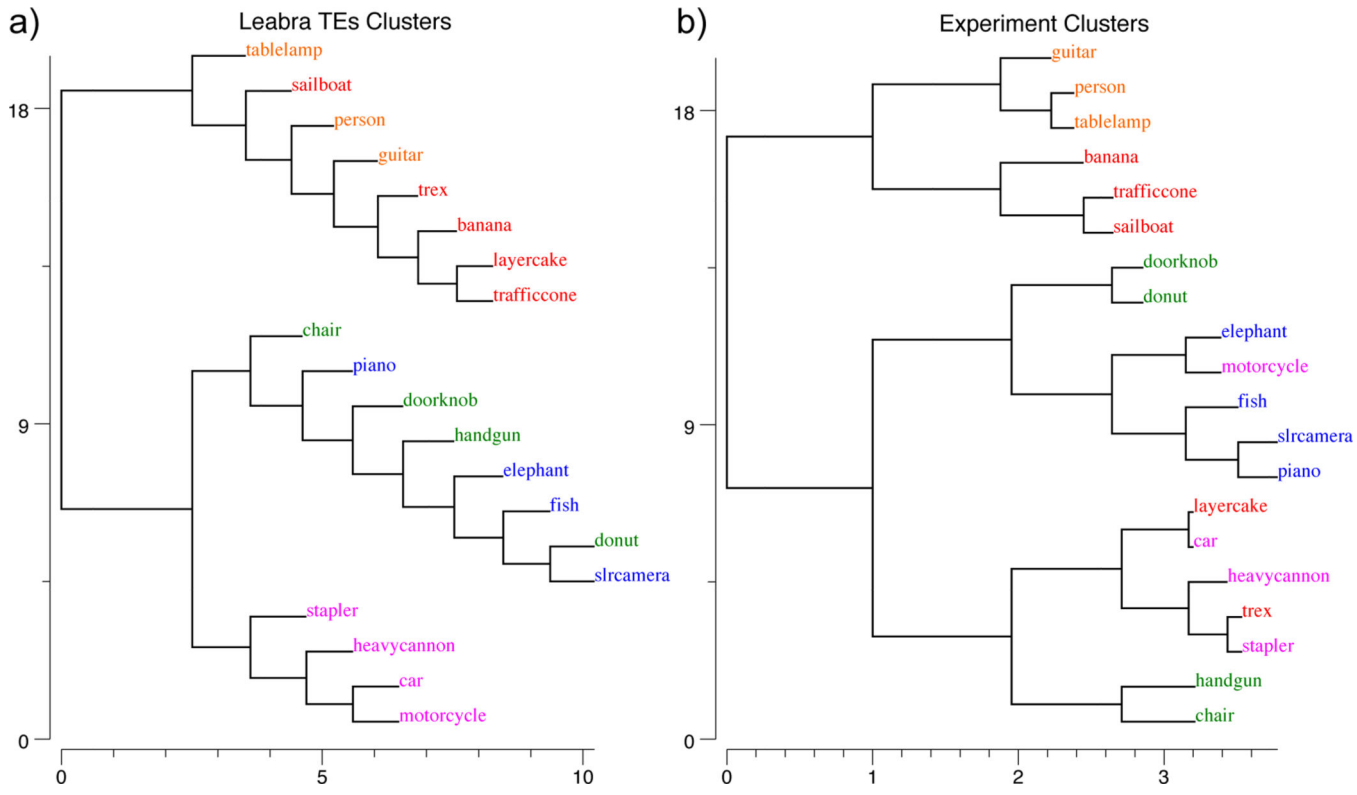

Figure 15:

Agglomerative clustering on the Leabra and Expt representations, with the centroid categories color coded. The most reliable information from this is the leaf-level groupings, as the rest of the structure is indeterminante and history dependent in reducing higher-dimensional structure down to a 2D plot. Both cluster plots show a strong tendency to group

leaf items together in the same centroid categories, with a few exceptions in each case. Also, the Leabra plot nicely captures the broader 3-category structure evident in the similarity matrix plots, within which the 5 finer-grained centroid categories are organized. Overall, this provides further confirmation that the model and the human subjects are organizing the shapes in largely the same way.

Furthermore, once we had collected the human experimental data (*Expt*), it was clear that it strongly coincided with our original shape intuitions, and with the finer-grained 5 category centroid structure. Starting from the centroid categories, the maximal permutation made only 3 changes, moving trex (T-rex) and handgun into the horizontal category, and chair into the pyramid, going from a distance score of 0.3083 to 0.3225, which is a relatively small improvement. However, using the maximal *Expt* clusters directly on the Leabra model gives a lower *acd* measure of 0.3745 (compared to 0.5071 for centroid), so the centroid categories represent a good middle-ground between experiment and the model, and this strong shared similarity structure with near-optimal cluster structures confirms that the model and people are encoding largely the same information.

In contrast, if we organize the experiment similarity matrix using the Bp categories, it produces a very poor average contrast distance measure of 0.0643 (compared to 0.3083 for the centroid categories), strongly suggesting that people's shape representations are not compatible with that simple structure.

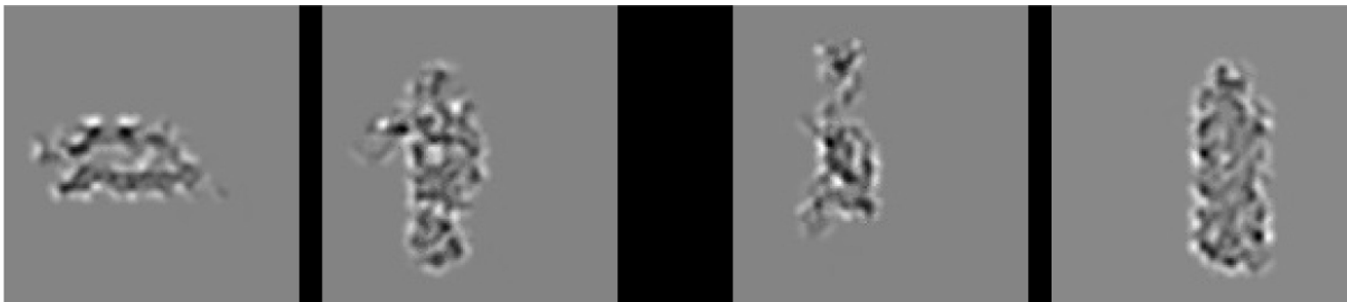

Figure 16:
Example stimulus from the behavioral experiment, using the V1 reconstruction of the actual input images presented to the model, to better capture the coarse-grained perception of the model. Subjects were requested to choose which of the two pairs, Left or Right, was most similar in terms of *overall shape*.

Another approach to determining clusters from similarity matrices, *agglomerative clustering*, starts with all items as singletons, and iteratively combines the closest two into a new cluster. The results for the Leabra and Expt similarity matrices are shown in Figure 15, which has also color-coded the items in terms of their category status according to the centroid structure. Due to a strong history dependency in the clustering process, and the indeterminacy of reducing a high-dimensional similarity structure down to two dimensions, structure beyond the leaf level is not very reliable (ties are also broken by a random number generator), but nevertheless you can clearly see that in both cases items from the same cluster are almost always together as leaves in the plots. This then provides additional

converging support for the idea that the model is learning the same kind of shape categories as people have.

For the network layer RSA computations, activation vectors were accumulated separately for each 3D object item, and within that separately for each frame index of the movie. To be able to monitor similarity metrics as the model trained, we used a running-average integration of neural activity across trials to accumulate the patterns. Specifically, the current activation pattern across each layer was recorded and averaged unit-by-unit with a time constant of $\tau = 10$. Critically, by integrating separately for each frame, this running-average computation did not introduce any bias for temporally-adjacent frames to be more similar. Nevertheless, when we computed the frame-to-frame similarities for TE, they were quite high (.901 correlation on average across all objects).

Behavioral Experiment Methods

The behavioral experiment was conducted on Amazon.com's MTurk web platform under University of Colorado IRB approval (19–0176), using 30 participants each categorizing up to 800 image pairs as shown in Figure 16, using the standard *simple image categorization* framework with a lightly customized script. Objects were drawn from the 156 3D object set, but data was aggregated in terms of the 20 basic-level categories (car, stapler, etc) because we could not sample all 156 × 156 object pairs. Thus, the resulting data was aggregated for each category pair in terms of the proportion of times when that pair was selected when presented.

The individual images were produced by reconstructing from the V1 transform that the computational model used in its high resolution V1 input layer, to give human participants as similar of an experience as possible to how the model "saw" the objects, and to reduce the influence of existing semantic knowledge which was entirely missing in our model (Figure 16).

Biological Model Methods

This section provides more information about the *DeepLeabra What-Where Integration (WWI)* model. The purpose of this information is to give more detailed insight into the model's function beyond the level provided in the main text, but with a model of this complexity, the only way to really understand it is to explore the model itself. It is available for download at: https://github.com/ccnlab/deep-obj-cat/tree/master/sims/cemer. We now have a full replication of this model in our new, much more transparent simulation framework, available at https://github.com/ccnlab/deep-obj-cat/tree/master/sims/wwi3d — this is more readable and recommended. Furthermore, the best way to understand this model is to understand the framework in which it is implemented, which is explained in great detail, with many running simulations explaining specific elements of functionality, at http://CompCogNeuro.org

Table 1:

Layer sizes, showing numbers of units in one pool (or entire layer if Pool is missing), and the number of Pools of such units, along X,Y axes. Each area has three associated layers: *s* = superficial layer, *d* = deep layer (context updated by 51B neurons in same area, shown in bold), *p* = pulvinar layer (driven by 5IB neurons from associated area, shown in bold).

		Units		Pools		
Area	**Name**	**X**	**Y**	**X**	**Y**	**Receiving Projections**
V1	V1s	4	5	8	8	
	V1p	4	5	8	8	V1s V2d V3d V4d TEOd
V1h	V1hs	4	5	16	16	
	V1hp	4	5	16	16	V1s V2d V3d V4d TEOd
Eyes	EyePos	21	21			
	SaccadePlan	11	11			
	Saccade	11	11			
Obj	ObjVel	11	11			
V2	V2s	10	10	8	8	V1s LIPs V3s V4s TEOd V1p V1hp
	V2d	10	10	8	8	**V2s** V1p V1hp LIPd LIPp V3d V4d V3s TEOs
LIP	MtPos	1	1	8	8	V1s
	LIPs	4	4	8	8	MtPos ObjVel SaccadePlan EyePos LIPp
	LIPd	4	4	8	8	**LIPs** LIPp ObjVel Saccade EyePos
	LIPp	1	1	8	8	**MtPos** V1s LIPd
V3	V3s	10	10	4	4	V2s V4s TEOs DPs LIPs V1p V1hp DPp TEOd
	V3d	10	10	4	4	**V3s** V1p V1hp DPp LIPd DPd V4d V4s DPs TEOs
	V3p	10	10	4	4	**V3s** V2d DPd TEOd
DP	DPs	10	10			V2s V3s TEOs V1p V1hp V3p TEOp
	DPd	10	10			**DPs** V1p V1hp DPp TEOd
	DPp	10	10			**DPs** V2d V3d DPd TEOd
V4	V4s	10	10	4	4	V2s TEOs V1p V1hp
	V4d	10	10	4	4	**V4s** V1p V1hp V4p TEOd TEOs
	V4p	10	10	4	4	**V4s** V2d V3d V4d TEOd
TEO	TEOs	10	10	4	4	V4s V1p V1hp TEs
	TEOd	10	10	4	4	**TEOs TEOd** V1p V1hp V4p TEOp TEp TEd
	TEOp	10	10	4	4	**TEOs** V3d V4d TEOd TEd
TE	TEs	10	10	4	4	TEOs V1p V1hp
	TEd	10	10	4	4	**TEs TEd** V1p V1hp V4p TEOp TEp TEOd
	TEp	10	10	4	4	**TEs** V3d V4d TEOd

Layer Sizes and Structure

Figure 5 in the main text shows the general configuration of the model, and Table 1 shows the specific sizes of each of the layers, and where they receive inputs from.

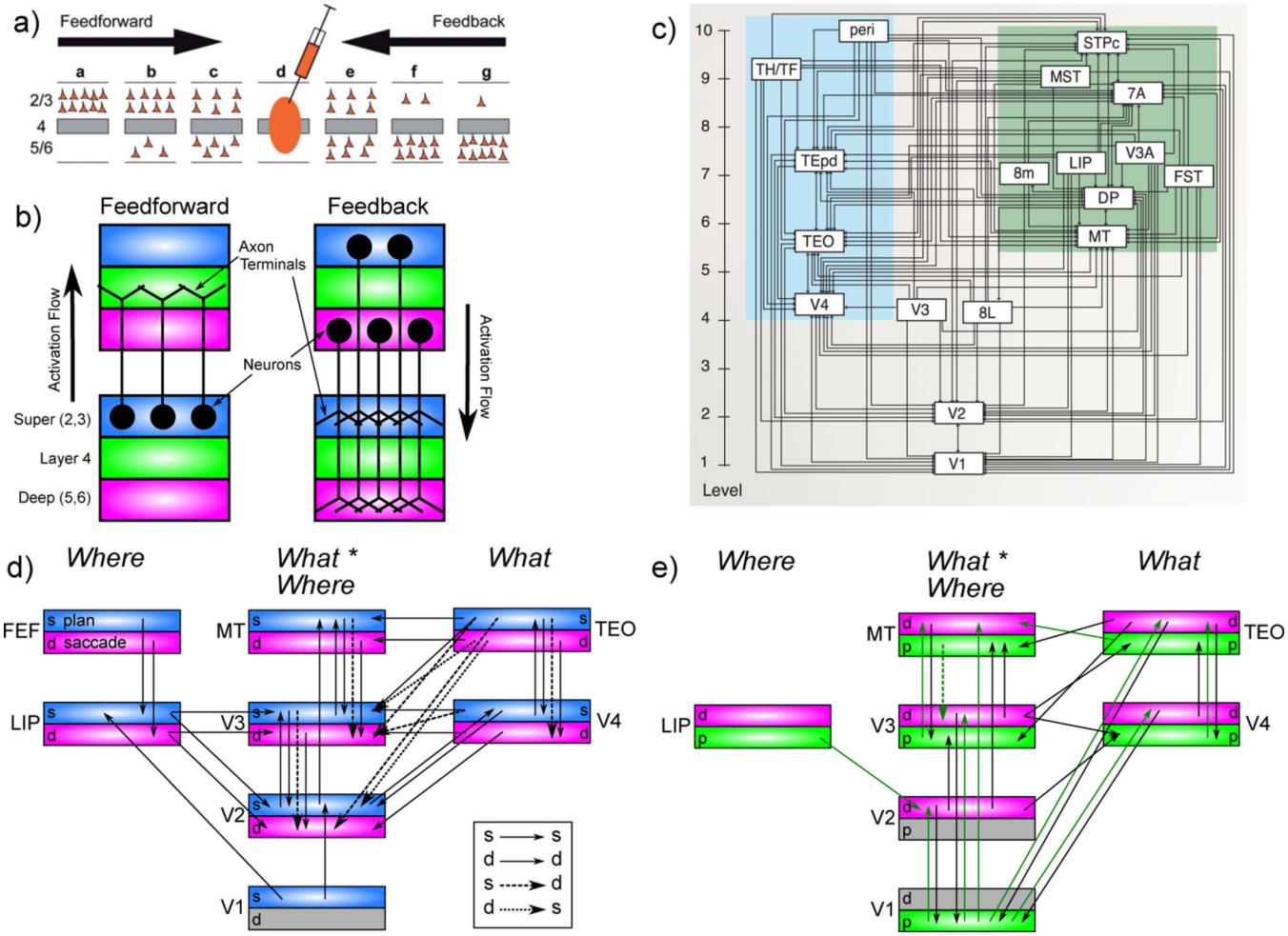

Figure 17:

Principles of connectivity in DeepLeabra. **a)** Markov et al (2014) data showing density of *retrograde* labeling from a given injection in a middle-level area (d): most feedforward projections originate from superficial layers of lower areas (a,b,c) and deep layers predominantly contribute to feedback (and more strongly for longer-range feedback). **b)** Summary diagram showing most feedforward connections originating in superficial layers of lower area, and terminating in layer 4 of higher area, while feedback connections can originate in either superficial or deep layers, and in both cases terminate in both superficial and deep layers of the lower area (adapted from Felleman & Van Essen, 1991). **c)** Anatomical hierarchy as determined by percentage of superficial layer source labeling (SLN) by Markov et al (2014) — the hierarchical levels are well matched for our model, but we functionally divide the dorsal pathway (shown in green background) into the two separable components of a *Where* and a *What * Where* integration pathway. **d)** Superficial and deep-layer connectivity in the model. Note the repeating motif between hierarchically-adjacent areas, with bidirectional connectivity between superficial layers, and feedback into deep layers from both higher-level superficial and deep layers, according to canonical pattern shown in panels a and b. Special patterns of connectivity from TEO to V3 and V2, involving crossed super-to-deep and deep-to-super pathways, provide top-down support

for predictions based on high-level object representations. **e)** Connectivity for deep layers and pulvinar in the model, which generally mirror the corticocortical pathways (in d). Each pulvinar layer (p) receives 5IB driving inputs from the labeled layer (e.g., V1p receives 5IB drivers from V1). In reality these neurons are more distributed throughout the pulvinar, but it is computationally convenient to organize them together as shown. Deep layers (d) provide predictive input into pulvinar, and pulvinar projections send error signals (via temporal differences between predictions and actual state) to *both* deep and superficial layers of given areas (only d shown). Most areas send deep-layer prediction inputs into the main V1p prediction layer, and receive reciprocal error signals therefrom. The strongest constraint we found was that pulvinar outputs (colored green) must generally project only to higher areas, not to lower areas, with the exceptions of DPp → V3 and LIPp → V2. V2p was omitted because it is largely redundant with V1p in this simple model.

All the activation and general learning parameters in the model are at their standard Leabra defaults.

Projections

The general principles and patterns of connectivity are shown in Figure 17 (and Figures 1 and 2 in the main text). As noted in the main text, the connectivity and overall structure obeys the established principles identified in neocortical anatomy (Felleman & Van Essen, 1991; Markov, Ercsey-Ravasz, et al., 2014; Markov, Vezoli, et al., 2014; Rockland & Pandya, 1979).

Detailing each of the specific parameters associated with the different projections shown in Table 1 would take too much space — those interested in this level of detail should download the model from the link shown above. There are topographic projections between many of the lower-level retinotopically-mapped layers, consistent with our earlier vision models (O'Reilly et al., 2013). For example the 8×8 unit groups in V2 are reduced down to the 4×4 groups in V3 via a 4×4 unit-group topographic projection, where neighboring units have half-overlapping receptive fields (i.e., the field moves over 2 unit groups in V2 for every 1 unit group in V3), and the full space is uniformly tiled by using a wrap-around effect at the edges. Similar patterns of connectivity are used in standard deep convolutional neural networks. However, we do not share weights across units as in a true convolutional network.

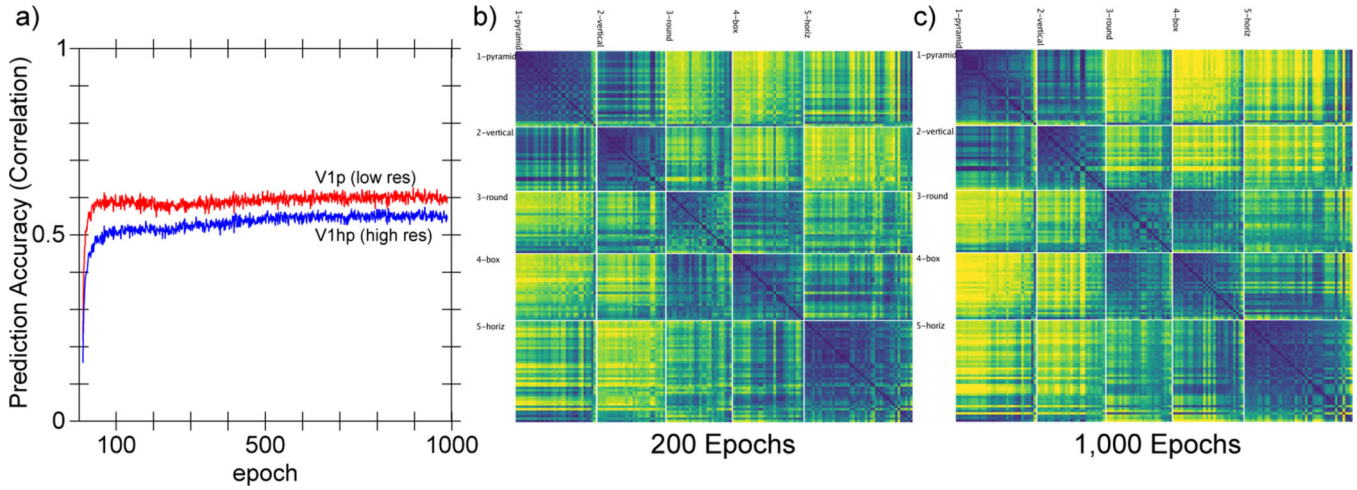

Figure 18:

a) Predictive learning curve for DeepLeabra, showing the correlation between prediction and actual over the two different V1 layers. Initial learning is quite rapid, followed by a slower but progressive learning process that reflects development of the IT representations (e.g., manipulations that interfere with those areas selectively impair this part of the learning curve). Overall prediction accuracy remains far from perfect, as shown in Figure 5 in main text, and significantly worse than the backpropagtionbased models. This is a typical finding from Leabra models which are significantly more constrained as a result of bidirectional attractor dynamics, Hebbian learning, and inhibitory competition – i.e., the very things that are likely important for forming abstract catgorical representations. **b)** Similarity matrix over TEs layer at 200 epochs, which has less contrast and definition (particularly evident in the off-block-diagonal differences) compared to the 1,000 epoch result (**c** also shown in Figure 7 in main text).

The projections from ObjVel (object velocity) and SaccadePlan layers to LIPs, LIPd were initialized with a topographic sigmoidal pattern that moved as a function of the position of the unit group, by a factor of .5, while the projections from EyePos were initialized with a gaussian pattern. These patterns multiplied uniformly distributed random weights in the .25 to .75 range, with the lowest values in the topographic pattern having a multiplier of .6, while the highest had a multiplier of 1 (i.e., a fairly subtle effect). This produced faster convergence of the LIP layer when doing *Where* pathway pre-training compared to purely random initial weights, consistent with Pouget and Sejnowski (1997) and related work on parietal gain field basis function representations.

In addition to exploring different patterns of overall connectivity, we also explored differences in the relative strengths of receiving projections, which can be set with a wt_scale.rel parameter in the simulator. All feedforward pathways have a default strength of 1. For the feedback projections, which are typically weaker (consistent with the biology), we explored a discrete range of strengths, typically .5, .2, .1, and .05. The strongest top-down projections were into V2s from LIP and V3, while most others were .2 or .1. Likewise projections from the pulvinar were weaker, typically .1. These differences in strength sometimes had large effects on performance during the initial bootstrapping of the overall

model structure, but in the final model they are typically not very consequential for any individual projection.

Training Parameters

Training typically consisted of 512 alpha trials per epoch (51.2 seconds of real time equivalent), for 1,000 such epochs. Each trial was generated from a virtual reality environment in the emergent simulator, that rendered first-person views with moving eye position onto the object tumbling through space with fixed motion and rotation parameters over the sequence of 8 frames (see Figure 5 in main text for representative example). Each frame was rendered at 256 × 256 resolution, and processed through our standard V1 gabor filters which are described in detail in O'Reilly et al. (2013).

Because the start of each sequence of 8 frames is unpredictable, we turned off learning for that trial, which improves learning overall. We have recently developed an automatic such mechanism based on the running-average (and running variance) of the prediction error, where we turn off learning whenever the current prediction error z-normalized by these running average values is below 1.5 standard deviations, which works well, and will be incorporated into future models. Biologically, this could correspond to a connection between pulvinar and neuromodulatory areas that could regulate the effective learning rate in this way.

Figure 18a shows the learning trajectory of the model, indicating that it learns quite rapidly. This rapid initial learning is likely facilitated by the extensive use of shortcut connections convering from all over the simulated visual system onto the V1 pulvinar layers, and direct projections back from these pulvinar layers. Thus, error signals are directly communicated and can drive learning quickly and efficiently. However, there are also extensive indirect, bidirectional connections among the superficial layers, which can drive indirect error backpropagation learning as well.

Model Algorithms

The biologically-based model was implemented using the Leabra framework, which is described in detail in previous publications (O'Reilly, 1996, 1998; O'Reilly et al., 2016; O'Reilly & Munakata, 2000; O'Reilly et al., 2012), and summarized here. The online textbook at https://CompCogNeuro.org provides the most comprehensive description of the framework, while https://github.com/emer/leabra has a summary of all the equations (and the code itself). There are two main implementations of Leabra, one in the *C++ emergent* software, and a new one using Go and Python language at the prior link. These same equations and standard parameters have been used to simulate over 40 different models in O'Reilly and Munakata (2000); O'Reilly et al. (2012), and a number of other research models. Thus, the model can be viewed as an instantiation of a systematic modeling framework using standardized mechanisms, instead of constructing new mechanisms for each model (O'Reilly et al., 2016).

The neurons use a rate code version of the adaptive exponential (AdEx) conductance-based point neuron model (Brette & Gerstner, 2005), with the standard RC circuit equations:

$$\Delta V_m(t) = \tau \sum_c g_c(t)\overline{g}_c(E_c - V_m(t)), \tag{2}$$

where c represents excitatory, inhibitory, and leak channels. Inhibition is driven by simulated interneurons in proportion to feedforward and feedback dynamics, producing sparse distributed representations, and controlling the effects of bidirectional excitatory connections between layers.

Each neuron learns using a more biologically-based version of the Contrastive Hebbian Learning (CHL) algorithm, as shown in Figure 2:

$$\Delta_{chl} = x^+y^+ - x^-y^- \tag{3}$$

where x is the sending activation, y is the receiving activation, and the + superscript indicates activations in the plus phase, and − those in the minus phase. The actual learning equations, detailed at https://github.com/emer/leabra and in the online textbook: https://CompCogNeuro.org produce a combination of error-driven and self-organizing factors, which emerge out of a single learning rule that was derived from a biologically detailed model of synaptic plasticity by (Urakubo et al., 2008), and is closely related to the Bienenstock, Cooper & Munro (BCM) algorithm (Bienenstock et al., 1982).

Deep Context

This section describes in detail the equations that are specific to the *Deep* version of Leabra that implements the specific predictive learning additions to the general algorithm. Like the simple recurrent network (SRN) (J. L. Elman, 1990; Jordan, 1989) which the deep predictive learning model functionally resembles, the primary computational specialization required is the maintenance of prior temporal context in the CT layer. In addition, the pulvinar layers have to be driven by the bottom-up inputs in the plus phase, after being driven by the CT inputs in the minus phase.

Computationally, the CT layer is specialized for maintaining context from the previous alpha cycle, to generate the prediction over the pulvinar layer. At the end of every plus phase, a new CT context excitatory input is computed from the normalized dot product of the context weights times the sending activations, just as in the standard net input used in Leabra:

$$\eta_j = \langle x_i w_{ij} \rangle = \frac{1}{n}\sum_i x_i w_{ij} \tag{4}$$

where x_i are the sending activations and w_{ij} are the weights. This net input is then added in with the standard net input at each cycle of processing during the subsequent alpha cycle.

The relative strength of these context layer inputs was set progressively larger for higher layers in the network, with a maximum of 4 in V4, TEO, and TE. In addition, TEO and TE received *self* context projections which provide an extended window of temporal context into the prior 200 ms interval, consistent with multiple sources of neural data (Chaudhuri et al., 2015). These self projections were connected only within the narrower Pool level

of units, enabling these neurons to develop mutually-excitatory loops to sustain activations over the multiple trials when the same object was present. We hypothesize that these modifications correspond to biological adaptations in IT cortex that likewise support greater sustained activation of object-level representations.

Learning of the context weights occurs as normal, but using the sending activation states from the *prior* time step's activation.

Computational and Biological Details of SRN-like Functionality

Predictive auto-encoder learning has been explored in various frameworks, but the most relevant to our model comes from the application of the SRN to a range of predictive learning domains (J. Elman et al., 1996; J. L. Elman, 1990). One of the most powerful features of the SRN is that it enables error-driven learning, instead of arbitrary parameter settings, to determine how prior information is integrated with new information. Thus, SRNs can learn to hold onto some important information for a relatively long interval, while rapidly updating other information that is only relevant for a shorter duration. This same flexibility is present in our DeepLeabra model. Furthermore, because this temporal context information is hypothesized to be present in the deep layers throughout the entire neocortex (in every microcolumn of tissue), the DeepLeabra model provides a more pervasive and interconnected form of temporal integration compared to the SRN, which typically just has a single temporal context layer associated with the internal "hidden" layer of processing units.

An extensive computational analysis of what makes the SRN work as well as it does, and explorations of a range of possible alternative frameworks, has led us to an important general principle: *subsequent outcomes determine what is relevant from the past*. At some level, this may seem obvious, but it has significant implications for predictive learning mechanisms based on temporal context. It means that the information encoded in a temporal context representation cannot be learned at the time when that information is presently active. Instead, the relevant contextual information is learned on the basis of what happens next.

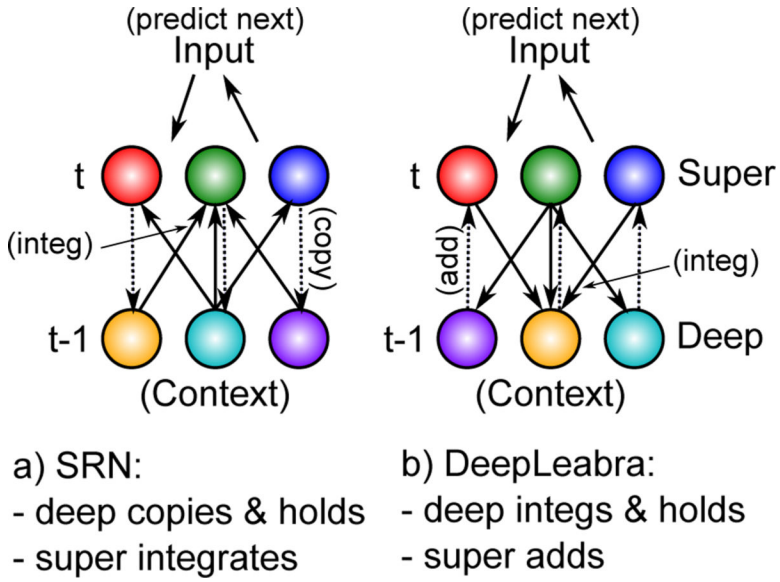

Figure 19:

How the DeepLeabra temporal context computation compares to the SRN mathematically. **a)** In a standard SRN, the context (deep layer biologically) is a copy of the hidden activations from the prior time step, and these are held constant while the hidden layer (superficial) units integrate the context through learned synaptic weights. **b)** In DeepLeabra, the deep layer performs the weighted integration of the soon-to-be context information from the superficial layer, and then holds this integrated value, and feeds it back as an additive net-input like signal to the superficial layer. The context net input is pre-computed, instead of having to compute this same value over and over again. This is more efficient, and more compatible with the diffuse interconnections among the deep layer neurons. Layer 6 projections to the thalamus and back recirculate this pre-computed net input value into the superficial layers (via layer 4), and back into itself to support maintenance of the held value.

This explains the peculiar power of the otherwise strange property of the SRN: the temporal context information is preserved as a *direct copy* of the state of the hidden layer units on the previous time step (Figure 19), and then learned synaptic weights integrate that copied context information into the next hidden state (which is then copied to the context again, and so on). This enables the error-driven learning taking place in the *current* time step to determine how context information from the *previous* time step is integrated. And the simple direct copy operation eschews any attempt to shape this temporal context itself, instead relying on the learning pressure that shapes the hidden layer representations to also shape the context representations. In other words, this copy operation is essential, because there is no other viable source of learning signals to shape the nature of the context representation itself (because these learning signals require future outcomes, which are by definition only available later).

The direct copy operation of the SRN is however seemingly problematic from a biological perspective: how could neurons copy activations from another set of neurons at some discrete point in time, and then hold onto those copied values for a duration of 100 ms,

which is a reasonably long period of time in neural terms (e.g., a rapidly firing cortical neuron fires at around 100 Hz, meaning that it will fire 10 times within that context frame). However, there is an important transformation of the SRN context computation, which is more biologically plausible, and compatible with the structure of the deep network (Figure 19). Specifically, instead of copying an entire set of activation states, the context activations (generated by the phasic 5IB burst) are immediately sent through the adaptive synaptic weights that integrate this information, which we think occurs in the 6CC (corticortical) and other lateral integrative connections from 5IB neurons into the rest of the deep network.

The result is a *pre-computed net input* from the context onto a given hidden unit (in the original SRN terminology), not the raw context information itself. Computationally, and metabolically, this is a much more efficient mechanism, because the context is, by definition, unchanging over the 100 ms alpha cycle, and thus it makes more sense to pre-compute the synaptic integration, rather than repeatedly re-computing this same synaptic integration over and over again (in the original feedforward backpropagation-based SRN model, this issue did not arise because a single step of activation updating took place for each context update — whereas in our bidirectional model many activation update steps must take place per context update).

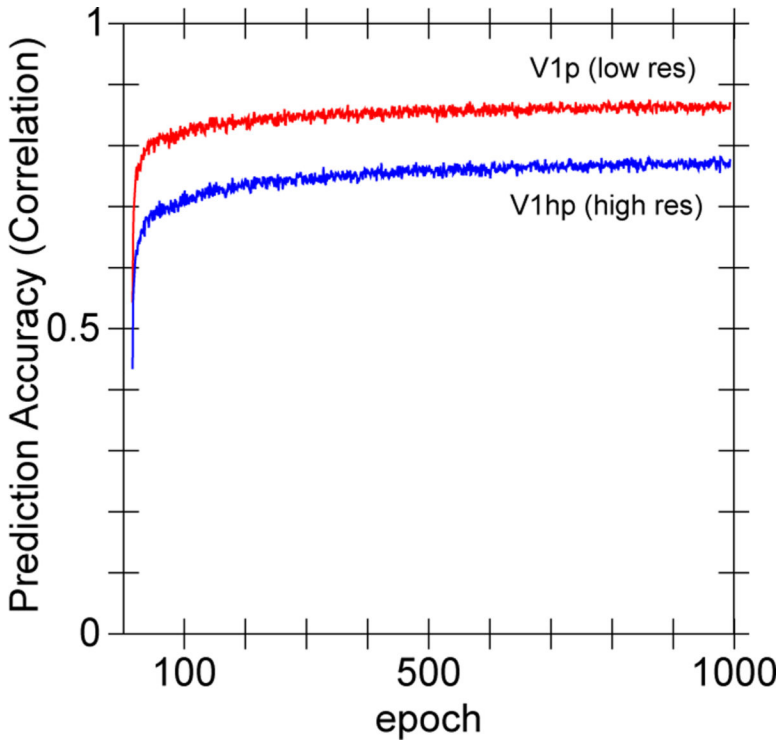

Figure 20:
Learning curves for the backpropagation version of the WWI model. Although it achieves better predictive accuracy than the DeepLeabra version, it fails to acquire abstract object category structure, indicating a potential tradeoff between simplifying and categorizing inputs, versus predicting precisely where the low-level visual features will move.

There are a couple of remaining challenges for this transformation of the SRN. First, the pre-computed net input from the context must somehow persist over the subsequent 100 ms period of the alpha cycle. We hypothesize that this can occur via NMDA and mGluR channels that can easily produce sustained excitatory currents over this time frame. Furthermore, the reciprocal excitatory connectivity from 6CT to TRC and back to 6CT could help to sustain the initial temporal context signal. Second, these contextual integration synapses require a different form of learning algorithm that uses the sending activation from the prior 100 ms, which is well within the time constants in the relevant calcium and second messenger pathways involved in synaptic plasticity.

Backpropagation Model Methods

The backpropagation version of the WWI model has exactly the same layer sizes and *feedforward* patterns of connectivity as the DeepLeabra version. Topographically, the V1p and V1hp pulvinar layers serve as output layers at the highest level of the network, receiving all the various connections from deep layers as shown in Table 1. Likewise, the LIPp served as a target output layer for the Where pathway. To achieve predictive learning, the V1 pulvinar targets were from the scene at time t, while the V1s inputs were from the scene at time $t - 1$. We also ran a comparison auto-encoder model that had inputs and target outputs from the same time step, and it showed even less systematic organization of its higher-level representations, further supporting the notion that predictive learning is important, across all frameworks. The learning curve for the predictive version is shown in Figure 20, which shows better overall prediction accuracy compared to the DeepLeabra model. However, as the RSA showed, this backpropagation model failed to learn object categories that go beyond the input similarity structure, indicating that perhaps it was paying too much "attention" in learning to this low-level structure, and lacked the necessary mechanisms to enable it to impose a simplifying higher-level structure on top of these inputs.

PredNet Model Methods

The PredNet architecture was designed to incorporate principles from predictive coding theory into a neural network model for predicting the next frame in a video sequence. Details of the model can be found in the original paper (Lotter et al., 2016), but here we provide a brief overview of the architecture.

Architecture

PredNet is a deep convolutional neural network that is composed of layers containing discrete modules. The lowest layer generates a prediction of incoming inputs (i.e. the pixels in the next frame), while each of the higher layers attempts to predict the *errors* made by the previous layer. Each layer contains an input convolutional module (A_l), a recurrent representational module (R_l), a prediction module $\left(\hat{A}_l\right)$, and a representation of its own errors (E_l). The input convolutional module (A_l) transforms its input with a set of standard convolutional filters, a rectified linear activation function, and a max-pooling operation. The recurrent representation module (R_l) is a convolutional LSTM, which is a recurrent convolutional network that replaces the matrix multiplications in the standard LSTM

equations with convolutions, allowing it to maintain a spatially organized representation of its inputs over time. The prediction module $(\hat{A}_l)$ consists of another standard convolutional layer and rectified linear activation that is used to generate predictions from the output of R_l. These predictions are then compared against the output of the input convolutional module (A_l). The errors generated in this comparison are represented explicitly in E_l, which applies a rectified linear activation to a concatenation of the positive $(A_l - \hat{A}_l)$ and negative $(\hat{A}_l - A_l)$ prediction errors. These errors then become the inputs to the next layer.

$$A_l^t = \begin{cases} x_t, & \text{if } l = 0 \\ MaxPool(ReLU(Conv(E_{l-1}^t))), & \text{if } l > 0 \end{cases} \quad (5)$$

$$\hat{A}_l^t = ReLU(Conv(R_l^t)) \quad (6)$$

$$E_l^t = \left[ReLU\left(A_l^t - \hat{A}_l^t\right); ReLU\left(\hat{A}_l^t - A_l^t\right)\right] \quad (7)$$

$$R_l^t = ConvLSTM\left(E_l^{t-1}, R_l^{t-1}, UpSample\left(R_{l+1}^t\right)\right) \quad (8)$$

At each time step in the video sequence, PredNet generates a prediction of the next frame. This is done as follows: first, the R_l is computed for each layer starting from the top of the hierarchy (because each R_l^t depends on input from R_{l+1}^t), and then the A_l^t, $\hat{A}_l^t$ and E_l^t are computed in a feed-forward fashion (becauase each A_l^t depends on input from the layer below, E_{l-1}^t).

All analyses in the RSA were conducted using the representations from the R_l layers.

Implementation details

All experiments with the PredNet architecture were performed using PyTorch. An informal hyperparameter search was conducted to find the settings that maximized representational similarity to the human judgments. This was done by conducting RSA on each layer for each hyperparameter setting, and computing, according to the Centroid categories derived from the human data, the difference between the average within-category similarity and the average between-category similarity. Our final architecture had 6 layers with 3, 16, 32, 64, 128, and 256 filters in the A_l and R_l modules, and 3×3 kernels throughout the whole network. We also found that using sigmoid and tanh activation functions in fully-connected convolutional LSTMs slightly improved performance, so these were used for all experiments.

The weights in the PredNet model are trained using error backpropagation. Predictions are generated and errors are computed at all levels of the hierarchy, but the model performs better when only the lowest layer's errors are backpropagated (Lotter et al., 2016). We confirmed these results with experiments that backpropagated the errors in higher layers, in which performance (in terms of mean squared error) was marginally reduced but the

RSA results were similar. For this reason, all reported experiments used a PredNet that was trained by only backpropagating the lowest level error.

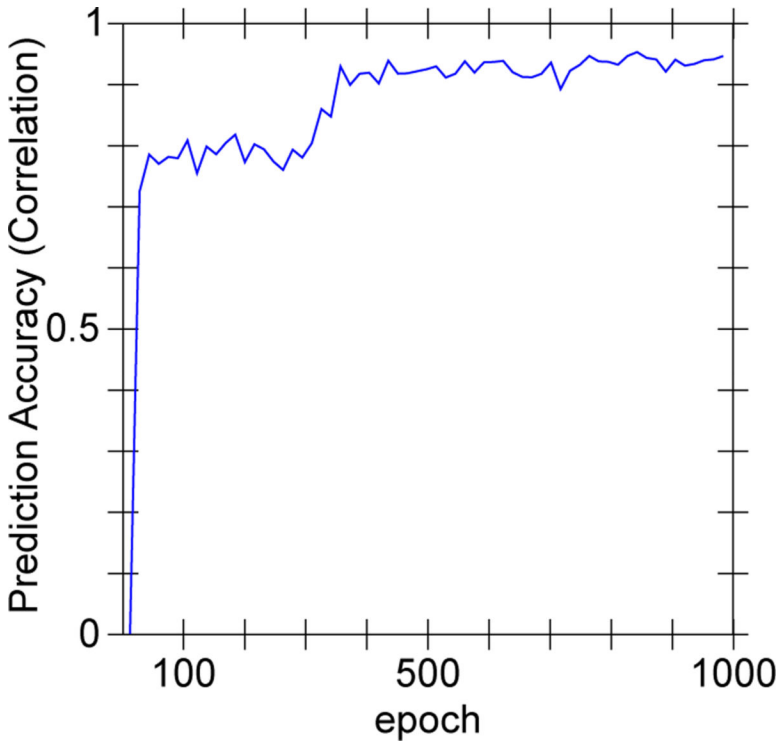

Figure 21:
Learning curves for the PredNet model. This model achieves the best overall prediction performance but also has the least well differentiated, categorical representations.

The model was trained using a batch size of 8 and an Adam optimizer with a learning rate of 0.0001, with no scheduler, for 150,000 batches. A training curve is shown in Figure 21, showing that it achieves the best overall prediction accuracy of any model we tested, and yet does not have representations that are as differentiated or categorical as our biologically based model, as shown in the main paper.

Regularization experiments

As discussed in the main paper, our biologically based model includes a number of important biologically motivated properties that may be contributing to the development of its categorical representations. These properties, including excitatory bidirectional connections, inhibitory competition, and an additional form of Hebbian learning, may be acting as regularizers that encourage categorical learning. We therefore tested whether standard regularization methods used in deep learning would have similar effects on the representations developed in the PredNet architecture. We tested 1) batch normalization, 2) dropout (0.1, 0.3, and 0.5), and 3) weight decay (0.01,0.001,0.0001,0.00001). All experiments with batch normalization and weight decay showed reduced performance (in terms of both prediction error on the test set and within-category correlation). As shown in figure 22, dropout marginally improved the within-category correlation while also slightly

improving prediction accuracy, so a dropout rate of 0.1 was used for the comparison to our biologically based model in the main paper.

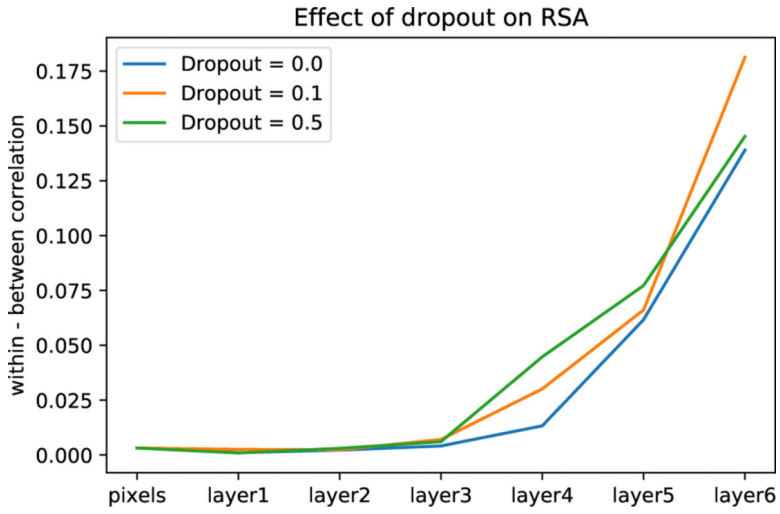

Figure 22:

Effect of dropout in PredNet on RSA, as measured by the difference between the average within-category correlation and the average between category correlation (using the Centroid categories derived from human data). Dropout marginally improves the category structure learned in PredNet.

References

Abbott LF, Varela JA, Sen K, & Nelson SB (1997, December). Synaptic depression and cortical gain control. Science, 275, 220. [PubMed: 8985017]

Ackley DH, Hinton GE, & Sejnowski TJ (1985, December). A learning algorithm for Boltzmann machines. Cognitive Science, 9(1), 147–169.

Antonov PA, Chakravarthi R, & Andersen SK (2020, October). Too little, too late, and in the wrong place: Alpha band activity does not reflect an active mechanism of selective attention. NeuroImage, 219, 117006. doi: 10.1016/j.neuroimage.2020.117006

Arcaro MJ, Pinsk MA, & Kastner S. (2015, July). The anatomical and functional organization of the human visual pulvinar. Journal of Neuroscience, 35(27), 9848–9871. doi: 10.1523/JNEUROSCI.1575-14.2015 [PubMed: 26156987]

Ashby FG, & Maddox WT (2011, April). Human Category Learning 2.0. Annals of the New York Academy of Sciences, 1224, 147–161. doi: 10.1111/j.1749-6632.2010.05874.x [PubMed: 21182535]

Barczak A, O'Connell MN, McGinnis T, Ross D, Mowery T, Falchier A, & Lakatos P. (2018, August). Top-down, contextual entrainment of neuronal oscillations in the auditory thalamocortical circuit. Proceedings of the National Academy of Sciences, 115(32), E7605–E7614. doi: 10.1073/pnas.1714684115

Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, & Friston KJ (2012, November). Canonical microcircuits for predictive coding. Neuron, 76(4), 695–711. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23177956 [PubMed: 23177956]

Bastos AM, Vezoli J, Bosman CA, Schoffelen J-M, Oostenveld R, Dowdall JR, ... Fries P. (2015, January). Visual Areas Exert Feedforward and Feedback Influences through Distinct Frequency Channels. Neuron, 85(2), 390–401. doi: 10.1016/j.neuron.2014.12.018 [PubMed: 25556836]

Bednar JA (2012, September). Building a mechanistic model of the development and function of the primary visual cortex. Journal of physiology, Paris, 106(5–6). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22343520

Bender DB (1982, July). Receptive-field properties of neurons in the macaque inferior pulvinar. Journal of neurophysiology, 48. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7119838

Bender DB, & Youakim M. (2001, January). Effect of attentive fixation in macaque thalamus and cortex. Journal of neurophysiology, 85, 219–234. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11152722 [PubMed: 11152722]

Bengio Y, Mesnard T, Fischer A, Zhang S, & Wu Y. (2017, January). STDP-compatible approximation of backpropagation in an energy-based model. Neural Computation, 29(3), 555–577. doi: 10.1162/NECO_a_00934 [PubMed: 28095200]

Bengio Y, Yao L, Alain G, & Vincent P. (2013). Generalized Denoising Auto-Encoders as Generative Models. In Burges CJC, Bottou L, Welling M, Ghahramani Z, & Weinberger KQ (Eds.), Advances in Neural Information Processing Systems 26 (pp. 899–907). Curran Associates, Inc. Retrieved 2017–05-15, from http://papers.nips.cc/paper/5023-generalized-denoising-auto-encoders-as-generative-models.pdf

Berger H. (1929, December). Über das Elektrenkephalogramm des Menschen. Archiv für Psychiatrie und Nervenkrankheiten, 87(1), 527–570. doi: 10.1007/BF01797193

Bienenstock EL, Cooper LN, & Munro PW (1982, March). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. The Journal of Neuroscience, 2(2), 32–48. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7054394 [PubMed: 7054394]

Bjork RA (1994). Memory and metamemory considerations in the training of human beings. In Metacognition: Knowing about knowing (pp. 185–205). Cambridge, MA, US: The MIT Press.

Bortone DS, Olsen SR, & Scanziani M. (2014, April). Translaminar inhibitory cells recruited by layer 6 corticothalamic neurons suppress visual cortex. Neuron, 82. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24656931

Bourne JA, & Rosa MGP (2006, March). Hierarchical development of the primate visual cortex, as revealed by neurofilament immunoreactivity: Early maturation of the middle temporal area (MT). Cerebral Cortex, 16(3), 405–414. doi: 10.1093/cercor/bhi119 [PubMed: 15944371]

Brette R, & Gerstner W. (2005, November). Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. Journal of Neurophysiology, 94(5), 3637–3642. doi: 10.1152/jn.00686.2005 [PubMed: 16014787]

Bridge H, Leopold DA, & Bourne JA (2016, February). Adaptive Pulvinar Circuitry Supports Visual Cognition. Trends in Cognitive Sciences, 20(2), 146–157. doi: 10.1016/j.tics.2015.10.003 [PubMed: 26553222]

Buffalo EA, Fries P, Landman R, Buschman TJ, & Desimone R. (2011, July). Laminar differences in gamma and alpha coherence in the ventral stream. Proceedings of the National Academy of Sciences of the United States of America, 108(27), 11262–11267. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21690410 [PubMed: 21690410]

Busch NA, Dubois J, & VanRullen R. (2009, June). The phase of ongoing EEG oscillations predicts visual perception. The Journal of Neuroscience, 29(24), 7869–7876. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19535598 [PubMed: 19535598]

Buzsáki G. (2005). Theta rhythm of navigation: Link between path integration and landmark navigation, episodic and semantic memory. Hippocampus, 15(7), 827–840. doi: 10.1002/hipo.20113 [PubMed: 16149082]

Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, ... DiCarlo JJ (2014, December). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. PLoS Computational Biology, 10(12), e1003963. doi: 10.1371/journal.pcbi.1003963

Cavanagh P, Hunt AR, Afraz A, & Rolfs M. (2010, April). Visual stability based on remapping of attention pointers. Trends in Cognitive Sciences, 14(4), 147–153. doi: 10.1016/j.tics.2010.01.007 [PubMed: 20189870]

Chaudhuri R, Knoblauch K, Gariel M-A, Kennedy H, & Wang X-J (2015, October). A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex. Neuron, 88(2), 419–431. doi: 10.1016/j.neuron.2015.09.008 [PubMed: 26439530]

Clark A. (2013, June). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3), 181–204. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23663408 [PubMed: 23663408]

Clayton MS, Yeung N, & Kadosh RC (2018). The many characters of visual alpha oscillations. European Journal of Neuroscience, 48(7), 2498–2508. doi: 10.1111/ejn.13747 [PubMed: 29044823]

Cleeremans A, & McClelland JL (1991, January). Learning the structure of event sequences. Journal of Experimental Psychology: General, 120, 235–253. [PubMed: 1836490]

Colby CL, Duhamel JR, & Goldberg ME (1997, March). Visual, presaccadic, and cognitive activation of single neurons in monkey lateral intraparietal area. Journal of neurophysiology, 76, 2841. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8930237

Connors BW, Gutnick MJ, & Prince DA (1982, December). Electrophysiological properties of neocortical neurons in vitro. Journal of Neurophysiology, 48(6), 1302–1320. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6296328 [PubMed: 6296328]

Cooper LN, & Bear MF (2012, November). The BCM theory of synapse modification at 30: Interaction of theory with experiment. Nature Reviews Neuroscience, 13(11), 798–810. doi: 10.1038/nrn3353 [PubMed: 23080416]

Crick F. (1984, July). Function of the thalamic reticular complex: The searchlight hypothesis. Proceedings of the National Academy of Sciences of the United States of America, 81, 4586–4590. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6589612 [PubMed: 6589612]

Crick F. (1989, February). The recent excitement about neural networks. Nature, 337, 129–132. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2911347 [PubMed: 2911347]

Dayan P. (1993, January). Improving generalization for temporal difference learning: The successor representation. Neural Computation, 5(4), 613–624. Retrieved from http://cognet.mit.edu/journal/10.1162/neco.1993.5.4.613

Dayan P, Hinton GE, Neal RN, & Zemel RS (1995, January). The Helmholtz machine. Neural Computation, 7(5), 889–904. [PubMed: 7584891]

de Lange FP, Heilbron M, & Kok P. (2018, September). How do expectations shape perception? Trends in Cognitive Sciences, 22(9), 764–779. doi: 10.1016/j.tics.2018.06.002 [PubMed: 30122170]

Desimone R, & Duncan J. (1995). Neural mechanisms of selective visual attention. Annual Review of Neuroscience, 18(1), 193–222. doi: 10.1146/annurev.ne.18.030195.001205

Duhamel JR, Colby CL, & Goldberg ME (1992, April). The updating of the representation of visual space in parietal cortex by intended eye movements. Science, 255(5040), 90–92. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1553535 [PubMed: 1553535]

Elman J, Bates E, Karmiloff-Smith A, Johnson M, Parisi D, & Plunkett K. (1996). Rethinking Innateness: A Connectionist Perspective on Development. Cambridge, MA: MIT Press.

Elman JL (1990, January). Finding structure in time. Cognitive Science, 14(2), 179–211.

Felleman DJ, & Van Essen DC (1991, January). Distributed Hierarchical Processing in the Primate Cerebral Cortex. Cerebral Cortex, 1(1), 1–47. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1822724 [PubMed: 1822724]

Fiebelkorn IC, & Kastner S. (2019, February). A rhythmic theory of attention. Trends in Cognitive Sciences, 23(2), 87–101. doi: 10.1016/j.tics.2018.11.009 [PubMed: 30591373]

Fiebelkorn IC, Pinsk MA, & Kastner S. (2018, August). A dynamic interplay within the frontoparietal network underlies rhythmic spatial attention. Neuron, 99(4), 842–853.e8. doi: 10.1016/j.neuron.2018.07.038 [PubMed: 30138590]

Fiser A, Mahringer D, Oyibo HK, Petersen AV, Leinweber M, & Keller GB (2016, December). Experience-dependent spatial expectations in mouse visual cortex. Nature Neuroscience, 19(12), 1658–1664. doi: 10.1038/nn.4385 [PubMed: 27618309]

Foldiak P. (1991, January). Learning Invariance from Transformation Sequences. Neural Computation, 3(2), 194–200. [PubMed: 31167302]

Foster JJ, & Awh E. (2019, October). The role of alpha oscillations in spatial attention: Limited evidence for a suppression account. Current Opinion in Psychology, 29, 34–40. doi: 10.1016/j.copsyc.2018.11.001 [PubMed: 30472541]

Franceschetti S, Guatteo E, Panzica F, Sancini G, Wanke E, & Avanzini G. (1995, October). Ionic mechanisms underlying burst firing in pyramidal neurons: Intracellular study in rat sensorimotor cortex. Brain Research, 696(1–2), 127–139. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8574660 [PubMed: 8574660]

Fries P, Womelsdorf T, Oostenveld R, & Desimone R. (2008, April). The Effects of Visual Stimulation and Selective Visual Attention on Rhythmic Neuronal Synchronization in Macaque Area V4. Journal of Neuroscience, 28(18), 4823–4835. doi: 10.1523/JNEUROSCI.4499-07.2008 [PubMed: 18448659]

Friston K. (2005, April). A theory of cortical responses. Philosophical Transactions of the Royal Society B, 360(1456), 815–836. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15937014

Friston K. (2010, February). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20068583 [PubMed: 20068583]

Fusi S, Miller EK, & Rigotti M. (2016, April). Why neurons mix: High dimensionality for higher cognition. Current Opinion in Neurobiology, 37, 66–74. doi: 10.1016/j.conb.2016.01.010 [PubMed: 26851755]

Gardner MPH, Schoenbaum G, & Gershman SJ (2018, November). Rethinking dopamine as generalized prediction error. Proceedings of the Royal Society B: Biological Sciences, 285(1891), 20181645. doi: 10.1098/rspb.2018.1645

Gavornik JP, & Bear MF (2014, May). Learned spatiotemporal sequence recognition and prediction in primary visual cortex. Nature Neuroscience, 17(5), 732–737. doi: 10.1038/nn.3683 [PubMed: 24657967]

George D, & Hawkins J. (2009, October). Towards a mathematical theory of cortical micro-circuits. PLoS Computational Biology, 5(10). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19816557

Goodale MA, & Milner AD (1992, January). Separate visual pathways for perception and action. Trends in Neurosciences, 15(1), 20–25. [PubMed: 1374953]

Gottlieb JP, Kusunoki M, & Goldberg ME (1998, February). The representation of visual salience in monkey parietal cortex. Nature, 391, 481. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9461214 [PubMed: 9461214]

Grill-Spector K, Henson R, & Martin A. (2006, January). Repetition and the brain: Neural models of stimulus-specific effects. Trends in Cognitive Sciences, 10(1), 14–23. doi: 10.1016/j.tics.2005.11.006 [PubMed: 16321563]

Grossberg S. (1999). How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex. Spatial vision, 12. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10221426

Gruber WR, Klimesch W, Sauseng P, & Doppelmayr M. (2005, April). Alpha Phase Synchronization Predicts P1 and N1 Latency and Amplitude Size. Cerebral Cortex, 15(4), 371–377. doi: 10.1093/cercor/bhh139 [PubMed: 15749980]

Gundlach C, Moratti S, Forschack N, & Muller MM (2020, May). Spatial Attentional Selection Modulates Early Visual Stimulus Processing Independently of Visual Alpha Modulations. Cerebral Cortex, 30(6), 3686–3703. doi: 10.1093/cercor/bhz335 [PubMed: 31907512]

Halassa MM, & Kastner S. (2017, December). Thalamic functions in distributed cognitive control. Nature Neuroscience, 20(12), 1669. doi: 10.1038/s41593-017-0020-1 [PubMed: 29184210]

Harris KD, & Shepherd GMG (2015, February). The neocortical circuit: Themes and variations. Nature Neuroscience, 18(2), 170–181. doi: 10.1038/nn.3917 [PubMed: 25622573]

Hawkins J, & Blakeslee S. (2004). On Intelligence. New York, NY: Times Books.

Hennig MH (2013). Theoretical models of synaptic short term plasticity. Frontiers in Computational Neuroscience, 7. Retrieved from http://www.frontiersin.org/computational_neuroscience/10.3389/fncom.2013.00045/abstract

Hinton GE, & McClelland JL (1988, January). Learning representations by recirculation. In Anderson DZ (Ed.), Neural Information Processing Systems (NIPS 1987) (Vol. 0, pp. 358–366). New York: American Institute of Physics. Retrieved from http://papers.nips.cc/paper/78-learning-representations-by-recirculation.pdf

Hinton GE, & Salakhutdinov RR (2006, July). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504–507. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16873662 [PubMed: 16873662]

Holroyd CB, & Coles MGH (2002, October). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. Psychological Review, 109(4), 679–709. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12374324 [PubMed: 12374324]

Hopfield JJ (1984, July). Neurons with graded response have collective computational properties like those of two-state neurons. Proceedings of the National Academy of Sciences USA, 81, 3088–3092. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6587342

Issa EB, Cadieu CF, & DiCarlo JJ (2018, November). Neural dynamics at successive stages of the ventral visual stream are consistent with hierarchical error signals. eLife, 7, e42870. doi: 10.7554/eLife.42870

Jaegle A, & Ro T. (2013, October). Direct Control of Visual Perception with Phase-specific Modulation of Posterior Parietal Cortex. Journal of Cognitive Neuroscience, 26(2), 422–432. doi: 10.1162/jocna00494 [PubMed: 24116843]

Jaramillo J, Mejias JF, & Wang X-J (2019, January). Engagement of Pulvino-cortical Feedforward and Feedback Pathways in Cognitive Computations. Neuron, 101(2), 321–336.e9. doi: 10.1016/j.neuron.2018.11.023 [PubMed: 30553546]

Jensen O, Bonnefond M, Marshall TR, & Tiesinga P. (2015, April). Oscillatory mechanisms of feedforward and feedback visual processing. Trends in Neurosciences, 38(4), 192–194. doi:10.1016/j.tins.2015.02.006 [PubMed: 25765320]

Jensen O, Bonnefond M, & VanRullen R. (2012, April). An oscillatory mechanism for prioritizing salient unattended stimuli. Trends in Cognitive Sciences, 16(4), 200–206. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22436764 [PubMed: 22436764]

Jensen O, & Mazaheri A. (2010). Shaping functional architecture by oscillatory alpha activity: Gating by inhibition. Frontiers in Human Neuroscience, 4(186). doi: 10.3389/fnhum.2010.00186

Jordan MI (1989, January). Serial Order: A Parallel, Distributed Processing Approach. In Elman JL & Rumelhart DE (Eds.), Advances in Connectionist Theory: Speech. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kachergis G, Wyatte D, O'Reilly RC, de Kleijn R, & Hommel B. (2014, November). A continuous-time neural model for sequential action. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1655), 20130623. doi: 10.1098/rstb.2013.0623

Kahana MJ, Seelig D, & Madsen JR (2001, December). Theta returns. Current Opinion in Neurobiology, 11(6), 739–744. doi: 10.1016/s0959-4388(01)00278-1 [PubMed: 11741027]

Kawato M, Hayakawa H, & Inui T. (1993, January). A forward-inverse optics model of reciprocal connections between visual cortical areas. Network: Computation in Neural Systems, 4(4), 415–422. doi: 10.1088/0954-898X44001

Keitel C, Keitel A, Benwell CSY, Daube C, Thut G, & Gross J. (2019, April). Stimulus-Driven Brain Rhythms within the Alpha Band: The Attentional-Modulation Conundrum. Journal of Neuroscience, 39(16), 3119–3129. doi: 10.1523/JNEUROSCI.1633-18.2019 [PubMed: 30770401]

Kelly SP, Lalor EC, Reilly RB, & Foxe JJ (2006, June). Increases in Alpha Oscillatory Power Reflect an Active Retinotopic Mechanism for Distracter Suppression During Sustained Visuospatial Attention. Journal of Neurophysiology, 95(6), 3844–3851. doi: 10.1152/jn.01234.2005 [PubMed: 16571739]

Khaligh-Razavi S-M, & Kriegeskorte N. (2014, November). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLOS Computational Biology, 10(11), e1003915. doi: 10.1371/journal.pcbi.1003915

Kiorpes L, Price T, Hall-Haro C, & Anthony Movshon J. (2012, June). Development of sensitivity to global form and motion in macaque monkeys (Macaca nemestrina). Vision Research, 63, 34–42. doi: 10.1016/j.visres.2012.04.018 [PubMed: 22580018]

Klimesch W. (2011, August). Evoked alpha and early access to the knowledge system: The P1 inhibition timing hypothesis. Brain Research, 1408, 52–71. doi: 10.1016/j.brainres.2011.06.003 [PubMed: 21774917]

Klimesch W, Sauseng P, & Hanslmayr S. (2007, January). EEG alpha oscillations: The inhibition-timing hypothesis. Brain Research Reviews, 53(1), 63–88. doi: 10.1016/j.brainresrev.2006.06.003 [PubMed: 16887192]

Kobatake E, & Tanaka K. (1994, January). Neuronal selectivities to complex object features in the ventral visual pathway. Journal of Neurophysiology, 71(3), 856–867. [PubMed: 8201425]

Kogo N, & Trengove C. (2015). Is predictive coding theory articulated enough to be testable? Frontiers in Computational Neuroscience, 9. doi: 10.3389/fncom.2015.00111

Kok P, & de Lange FP (2015). Predictive Coding in Sensory Cortex. In An Introduction to Model-Based Cognitive Neuroscience (pp. 221–244). Springer, New York, NY. doi: 10.1007/978-1-4939-2236-9 11

Kok P, Jehee JFM, & de Lange FP (2012, July). Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. Neuron, 75(2), 265–270. doi: 10.1016/j.neuron.2012.04.034 [PubMed: 22841311]

Komura Y, Nikkuni A, Hirashima N, Uetake T, & Miyamoto A. (2013, June). Responses of pulvinar neurons reflect a subject's confidence in visual categorization. Nature Neuroscience, 16(6), 749–755. doi: 10.1038/nn.3393 [PubMed: 23666179]

Kriegeskorte N, Mur M, & Bandettini P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience, 2(4). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19104670

LaBerge D, & Buchsbaum MS (1990, March). Positron emission tomographic measurements of pulvinar activity during an attention task. The Journal of neuroscience : the official journal of the Society for Neuroscience, 10, 613–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2303863 [PubMed: 2303863]

Larkum ME, Zhu JJ, & Sakmann B. (1999, March). A new cellular mechanism for coupling inputs arriving at different cortical layers. Nature, 398(6725), 338–341. doi: 10.1038/18686 [PubMed: 10192334]

LeCun Y, Bengio Y, & Hinton G. (2015, May). Deep learning. Nature, 521(7553), 436–444. doi: 10.1038/nature14539 [PubMed: 26017442]

Lee TS, & Mumford D. (2003, July). Hierarchical Bayesian inference in the visual cortex. Journal of the Optical Society of America, 20(7), 1434–1448. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12868647/ [PubMed: 12868647]

Lillicrap TP, Santoro A, Marris L, Akerman CJ, & Hinton G. (2020, June). Backpropagation and the brain. Nature Reviews Neuroscience, 21(6), 335–346. doi: 10.1038/s41583-020-0277-3 [PubMed: 32303713]

Lim S, McKee JL, Woloszyn L, Amit Y, Freedman DJ, Sheinberg DL, & Brunel N. (2015, December). Inferring learning rules from distributions of firing rates in cortical neurons. Nature Neuroscience, 18(12), 1804–1810. doi: 10.1038/nn.4158 [PubMed: 26523643]

Lotter W, Kreiman G, & Cox D. (2016, May). Deep predictive coding networks for video prediction and unsupervised learning. arXiv:1605.08104 [cs, q-bio]. Retrieved 2017–08-11, from http://arxiv.org/abs/1605.08104

Luczak A, Bartho P, & Harris KD (2009, May). Spontaneous events outline the realm of possible sensory responses in neocortical populations. Neuron, 62(3), 413–425. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19447096 [PubMed: 19447096]

Luczak A, Bartho P, & Harris KD (2013, January). Gating of sensory input by spontaneous cortical activity. The Journal of Neuroscience, 33(4), 1684–1695. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23345241 [PubMed: 23345241]

Lüscher C, & Malenka RC (2012, June). NMDA receptor-dependent long-term potentiation and long-term depression (LTP/LTD). Cold Spring Harbor Perspectives in Biology, 4(6), a005710. doi: 10.1101/cshperspect.a005710

Maier A, Adams GK, Aura C, & Leopold DA (2010). Distinct Superficial and Deep Laminar Domains of Activity in the Visual Cortex during Rest and Stimulation. Frontiers in Systems Neuroscience, 4(31). doi: 10.3389/fnsys.2010.00031

Maier A, Aura CJ, & Leopold DA (2011, February). Infragranular sources of sustained local field potential responses in macaque primary visual cortex. The Journal of Neuroscience, 31(6), 1971–1980. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21307235 [PubMed: 21307235]

Makeig S, Westerfield M, Jung TP, Enghoff S, Townsend J, Courchesne E, & Sejnowski TJ (2002, January). Dynamic Brain Sources of Visual Evoked Responses. Science, 295, 690–693. [PubMed: 11809976]

Marino AC, & Mazer JA (2016). Perisaccadic Updating of Visual Representations and Attentional States: Linking Behavior and Neurophysiology. Frontiers in Systems Neuroscience, 10. doi: 10.3389/fnsys.2016.00003

Markov NT, Ercsey-Ravasz MM, Gomes R, R A, Lamy C, Magrou L, ... Kennedy H. (2014, January). A Weighted and Directed Interareal Connectivity Matrix for Macaque Cerebral Cortex. Cerebral Cortex, 24(1), 17–36. doi: 10.1093/cercor/bhs270 [PubMed: 23010748]

Markov NT, Vezoli J, Chameau P, Falchier A, Quilodran R, Huissoud C, ... Kennedy H. (2014, January). Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex: Cortical counterstreams. Journal of Comparative Neurology, 522(1), 225–259. doi: 10.1002/cne.23458 [PubMed: 23983048]

Martinez-Conde S, Macknik SL, & Hubel DH (2004, March). The role of fixational eye movements in visual perception. Nature Reviews Neuroscience, 5(3), 229–240. doi: 10.1038/nrn1348 [PubMed: 14976522]

Martinez-Conde S, Otero-Millan J, & Macknik SL (2013, February). The impact of microsaccades on vision: Towards a unified theory of saccadic function. Nature Reviews Neuroscience, 14(2), 83–96. doi: 10.1038/nrn3405 [PubMed: 23329159]

Mathewson K, Gratton G, Fabiani M, Beck D, & Ro T. (2009). To see or not to see: Prestimulus alpha phase predicts visual awareness. The Journal of Neuroscience, 29(9), 2725–2732. [PubMed: 19261866]

Mathewson KE, Fabiani M, Gratton G, Beck DM, & Lleras A. (2010, April). Rescuing stimuli from invisibility: Inducing a momentary release from visual masking with pre-target entrainment. Cognition, 115(1), 186–191. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20035933 [PubMed: 20035933]

Mathewson KE, Prudhomme C, Fabiani M, Beck DM, Lleras A, & Gratton G. (2012, August). Making waves in the stream of consciousness: Entraining oscillations in EEG alpha and fluctuations in visual awareness with rhythmic visual stimulation. Journal of Cognitive Neuroscience, 24(12), 2321–2333. doi: 10.1162/jocna00288 [PubMed: 22905825]

Mayer A, Schwiedrzik CM, Wibral M, Singer W, & Melloni L. (2016, July). Expecting to See a Letter: Alpha Oscillations as Carriers of Top-Down Sensory Predictions. Cerebral Cortex, 26(7), 3146–3160. doi: 10.1093/cercor/bhv146 [PubMed: 26142463]

Meyer T, & Olson CR (2011, November). Statistical learning of visual transitions in monkey inferotemporal cortex. Proceedings of the National Academy of Sciences of the United States of America, 108(48), 19401–19406. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22084090 [PubMed: 22084090]

Michalareas G, Vezoli J, van Pelt S, Schoffelen J-M, Kennedy H, & Fries P. (2016, January). Alpha-Beta and Gamma Rhythms Subserve Feedback and Feedforward Influences among Human Visual Cortical Areas. Neuron, 89(2), 384–397. doi: 10.1016/j.neuron.2015.12.018 [PubMed: 26777277]

Miller EK, & Cohen JD (2001). An integrative theory of prefrontal cortex function. Annual Review of Neuroscience, 24, 167–202. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11283309

Miller KD (1994, February). 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. The Journal of Neuroscience, 14(1), 409–441. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8283248 [PubMed: 8283248]

Müller JR, Metha AB, Krauskopf J, & Lennie P. (1999, September). Rapid adaptation in visual cortex to the structure of images. Science (New York, N.Y.), 285, 1405. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10464100 [PubMed: 10464100]

Mumford D. (1991, June). On the computational architecture of the neocortex. Biological Cybernetics, 65(2), 135–145. doi: 10.1007/BF00202389 [PubMed: 1912004]

Mumford D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biological Cybernetics, 66(3), 241–251. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1540675 [PubMed: 1540675]

Nakamura K, & Colby CL (2002, March). Updating of the visual representation in monkey striate and extrastriate cortex during saccades. Proceedings of the National Academy of Sciences of the United States of America, 99(6), 4026–4031. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11904446 [PubMed: 11904446]

Neupane S, Guitton D, & Pack CC (2016, February). Two distinct types of remapping in primate cortical area V4. Nature Communications, 7, 10402. doi: 10.1038/ncomms10402

Neupane S, Guitton D, & Pack CC (2017, July). Coherent alpha oscillations link current and future receptive fields during saccades. Proceedings of the National Academy of Sciences, 201701672. doi: 10.1073/pnas.1701672114

Neupane S, Guitton D, & Pack CC (2020, April). Perisaccadic remapping: What? How? Why? Reviews in the Neurosciences. doi: 10.1515/revneuro-2019-0097

Nunn CMH, & Osselton JW (1974, May). The Influence of the EEG Alpha Rhythm on the Perception of Visual Stimuli. Psychophysiology, 11(3), 294–303. doi: 10.1111/j.1469-8986.1974.tb00547.x [PubMed: 4421317]

O'Herron P, & von der Heydt R. (2013, January). Remapping of border ownership in the visual cortex. Journal of Neuroscience, 33(5), 1964–1974. doi: 10.1523/JNEUROSCI.2797-12.2013 [PubMed: 23365235]

Olsen S, Bortone D, Adesnik H, & Scanziani M. (2012, February). Gain control by layer six in cortical circuits of vision. Nature, 483(7387), 47–52. [PubMed: 22367547]

O'Reilly RC (1996, January). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. Neural Computation, 8(5), 895–938. doi: 10.1162/neco.1996.8.5.895

O'Reilly RC (1998, January). Six Principles for Biologically-Based Computational Models of Cortical Cognition. Trends in Cognitive Sciences, 2(11), 455–462. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/21227277 [PubMed: 21227277]

O'Reilly RC, Hazy TE, & Herd SA (2016). The Leabra cognitive architecture: How to play 20 principles with nature and win! In Chipman S. (Ed.), Oxford handbook of cognitive science. Oxford University Press. Retrieved 2015–05-15, from http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199842193.001.0001/oxfordhb-9780199842193-e-8

O'Reilly RC, & Munakata Y. (2000). Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain. Cambridge, MA: MIT Press.

O'Reilly RC, Munakata Y, Frank MJ, Hazy TE, & Contributors. (2012). Computational Cognitive Neuroscience. Wiki Book, 1st Edition, URL: http://ccnbook.colorado.edu. Retrieved from http://ccnbook.colorado.edu

O'Reilly RC, Wyatte D, Herd S, Mingus B, & Jilk DJ (2013). Recurrent Processing during Object Recognition. Frontiers in Psychology, 4(124). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23554596

O'Reilly RC, Wyatte D, & Rohrlich J. (2014, July). Learning Through Time in the Thalamocortical Loops. arXiv:1407.3432 [q-bio]. Retrieved 2015–05-15, from http://arxiv.org/abs/1407.3432

O'Reilly RC, Wyatte DR, & Rohrlich J. (2017, September). Deep predictive learning: A comprehensive model of three visual streams. arXiv:1709.04654 [q-bio]. Retrieved 2017–09-15, from http://arxiv.org/abs/1709.04654

Ouden HEM, Kok P, & Lange FP (2012). How prediction errors shape perception, attention, and motivation. Frontiers in Psychology, 3(548). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23248610

Palva S, & Palva JM (2011). Functional roles of alpha-band phase synchronization in local and large-scale cortical networks. Frontiers in Psychology, 2(204), ePub only. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21922012

Pennartz CM, Dora S, Muckli L, & Lorteije JA (2019). Towards a Unified View on Pathways and Functions of Neural Recurrent Processing. Trends in Neurosciences.

Petersen SE, Robinson DL, & Keys W. (1985, October). Pulvinar nuclei of the behaving rhesus monkey: Visual responses and their modulation. Journal of neurophysiology, 54. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4067625

Petrof I, Viaene AN, & Sherman SM (2012, June). Two populations of corticothalamic and interareal corticocortical cells in the subgranular layers of the mouse primary sensory cortices. Journal of Comparative Neurology, 520(8), 1678–1686. doi: 10.1002/cne.23006 [PubMed: 22120996]

Pinault D. (2004, August). The thalamic reticular nucleus: Structure, function and concept. Brain research, 46. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15297152

Pineda FJ (1987, January). Generalization of Backpropagation to Recurrent Neural Networks. Physical Review Letters, 18, 2229–2232.

Pouget A, & Sejnowski TJ (1997, January). Spatial Transformations in the Parietal Cortex Using Basis Functions. Journal of Cognitive Neuroscience, 9(2), 222–237. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23962013 [PubMed: 23962013]

Privman E, Malach R, & Yeshurun Y. (2013, April). Modeling the electrical field created by mass neural activity. Neural Networks, 40, 44–51. doi: 10.1016/j.neunet.2013.01.004 [PubMed: 23391515]

Purushothaman G, Marion R, Li K, & Casagrande VA (2012, June). Gating and control of primary visual cortex by pulvinar. Nature Neuroscience, 15(6), 905–912. doi: 10.1038/nn.3106 [PubMed: 22561455]

Pylyshyn Z. (1989, June). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. Cognition, 32(1), 65–97. doi: 10.1016/0010-0277(89)90014-0 [PubMed: 2752706]

Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, & DiCarlo JJ (2018, February). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. bioRxiv, 240614. doi: 10.1101/240614

Rao RP, & Ballard DH (1999, January). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 2(1), 79–87. doi: 10.1038/4580 [PubMed: 10195184]

Ray S, & Maunsell JHR (2011, April). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLoS biology, 9(4), e1000610. doi: 10.1371/journal.pbio.1000610

Reber AS (1967, January). Implicit Learning of Artificial Grammars. Journal of Verbal Learning and Verbal Behavior, 6, 855–863.

Reynolds JH, Chelazzi L, & Desimone R. (1999, April). Competitive mechanisms subserve attention in macaque areas V2 and V4. The Journal of neuroscience : the official journal of the Society for Neuroscience, 19, 1736–1753. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10024360 [PubMed: 10024360]

Reynolds JH, & Heeger DJ (2009, January). The normalization model of attention. Neuron, 61(2), 168–185. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19186161/ [PubMed: 19186161]

Richter D, & de Lange FP (2019, August). Statistical learning attenuates visual activity only for attended stimuli. eLife, 8, e47869. doi: 10.7554/eLife.47869

Robinson DL (1993). Functional contributions of the primate pulvinar. Progress in brain research, 95. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8493346

Rockland KS (1996, October). Two types of corticopulvinar terminations: Round (type 2) and elongate (type 1). The Journal of comparative neurology, 368, 57–87. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8725294 [PubMed: 8725294]

Rockland KS (1998, January). Convergence and branching patterns of round, type 2 corticopulvinar axons. The Journal of Comparative Neurology, 390(4), 515–536. doi: 10.1002/(SICI)1096-9861(19980126)390:4h515::AID-CNE5i3.0.CO;2-3 [PubMed: 9450533]

Rockland KS, & Pandya DN (1979, December). Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. Brain Research, 179(1), 3–20. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/116716 [PubMed: 116716]

Rumelhart DE, Hinton GE, & Williams RJ (1986, January). Learning representations by backpropagating errors. Nature, 323(9), 533–536.

Rumelhart DE, & McClelland JL (1982, April). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. Psychological review, 89, 60–94. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7058229 [PubMed: 7058229]

Saalmann YB, & Kastner S. (2011, July). Cognitive and perceptual functions of the visual thalamus. Neuron, 71(2), 209–223. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21791281 [PubMed: 21791281]

Saalmann YB, Pinsk MA, Wang L, Li X, & Kastner S. (2012, August). The pulvinar regulates information transmission between cortical areas based on attention demands. Science, 337(6095), 753–756. doi: 10.1126/science.1223082 [PubMed: 22879517]

Sakata S, & Harris KD (2009, November). Laminar structure of spontaneous and sensory-evoked population activity in auditory cortex. Neuron, 64(3), 404–418. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19914188 [PubMed: 19914188]

Sakata S, & Harris KD (2012). Laminar-dependent effects of cortical state on auditory cortical spontaneous activity. Frontiers in neural circuits, 6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23267317

Samaha J, Bauer P, Cimaroli S, & Postle BR (2015, July). Top-down control of the phase of alpha-band oscillations as a mechanism for temporal prediction. Proceedings of the National Academy of Sciences USA, 112(27), 8439–8444. doi: 10.1073/pnas.1503686112

Sherman MT, Kanai R, Seth AK, & VanRullen R. (2016, April). Rhythmic influence of top–down perceptual priors in the phase of prestimulus occipital alpha oscillations. Journal of Cognitive Neuroscience, 28(9), 1318–1330. doi: 10.1162/jocna00973 [PubMed: 27082046]

Sherman SM (2014, May). The function of metabotropic glutamate receptors in thalamus and cortex. The Neuroscientist, 20(2), 146–149.

Sherman SM, & Guillery RW (2006). Exploring the Thalamus and Its Role in Cortical Function. Cambridge, MA: MIT Press. Retrieved from http://www.scholarpedia.org/article/Thalamus

Sherman SM, & Guillery RW (2011, September). Distinct functions for direct and transthalamic corticocortical connections. Journal of Neurophysiology, 106(3), 1068–1077. doi: 10.1152/jn.00429.2011 [PubMed: 21676936]

Sherman SM, & Guillery RW (2013). Functional Connections of Cortical Areas: A New View From the Thalamus. Cambridge, MA: MIT Press.

Shipp S. (2003, October). The functional logic of cortico-pulvinar connections. Philosophical Transactions of the Royal Society of London B, 358(1438), 1605–1624. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14561322

Shouval HZS, Bear MF, & Cooper LN (2002, August). A unified model of NMDA receptor-dependent bidirectional synaptic plasticity. Proceedings of the National Academy of Sciences USA, 99(16), 10831–10836. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12136127

Shrager J, & Johnson MH (1996, October). Dynamic Plasticity Influences the Emergence of Function in a Simple Cortical Array. Neural Networks, 9(7), 1119–1129. doi: 10.1016/0893-6080(96)00033-0 [PubMed: 12662587]

Silva LR, Amitai Y, & Connors BW (1991, January). Intrinsic oscillations of neocortex generated by layer 5 pyramidal neurons. Science, 251(4992), 432–435. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1824881 [PubMed: 1824881]

Snow JC, Allen HA, Rafal RD, & Humphreys GW (2009, March). Impaired attentional selection following lesions to human pulvinar: Evidence for homology between human and monkey. Proceedings of the National Academy of Sciences, 106(10), 4054–4059. doi: 10.1073/pnas.0810086106

Soĺıs-Vivanco R, Jensen O, & Bonnefond M. (2018, August). Top-Down Control of Alpha Phase Adjustment in Anticipation of Temporally Predictable Visual Stimuli. Journal of Cognitive Neuroscience, 30(8), 1157–1169. doi: 10.1162/jocna01280 [PubMed: 29762100]

Solomon EA, Kragel JE, Sperling MR, Sharan A, Worrell G, Kucewicz M, ... Kahana MJ (2017, November). Widespread theta synchrony and high-frequency desynchronization underlies enhanced cognition. Nature Communications, 8(1), 1704. doi: 10.1038/s41467-017-01763-2

Spaak E, Bonnefond M, Maier A, Leopold DA, & Jensen O. (2012, December). Layer-specific entrainment of gamma-band neural activity by the alpha rhythm in monkey visual cortex. Current Biology, 22(24), 2313–2318. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23159599 [PubMed: 23159599]

Spaak E, de Lange FP, & Jensen O. (2014, March). Local Entrainment of Alpha Oscillations by Visual Stimuli Causes Cyclic Modulation of Perception. Journal of Neuroscience, 34(10), 3536–3544. doi: 10.1523/JNEUROSCI.4385-13.2014 [PubMed: 24599454]

Spelke E, Breinlinger K, Macomber J, & Jacobson K. (1992, January). Origins of Knowledge. Psychological Review, 99(4), 605–632. [PubMed: 1454901]

Spratling MW (2008). Reconciling predictive coding and biased competition models of cortical function. Frontiers in Computational Neuroscience, 2(4), 1–8 (online). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18978957 [PubMed: 18946531]

Summerfield C, & de Lange FP (2014, November). Expectation in perceptual decision making: Neural and computational mechanisms. Nature Reviews Neuroscience, 15(11), 745–756. doi: 10.1038/nrn3838 [PubMed: 25315388]

Summerfield C, & Egner T. (2009, September). Expectation (and attention) in visual cognition. Trends in Cognitive Sciences, 13(9), 403–409. doi: 10.1016/j.tics.2009.06.003 [PubMed: 19716752]

Summerfield C, Trittschuh EH, Monti JM, Mesulam MM, & Egner T. (2008, September). Neural repetition suppression reflects fulfilled perceptual expectations. Nature Neuroscience, 11(9), 1004–1006. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19160497 [PubMed: 19160497]

Sutton RS, & Barto AG (1998). Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press. Retrieved from http://www.cs.ualberta.ca/sutton/book/ebook/the-book.html

Thomson AM (2010). Neocortical layer 6, a review. Frontiers in Neuroanatomy, 4(13). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20556241

Thomson AM, & Lamy C. (2007, November). Functional maps of neocortical local circuitry. Frontiers in Neuroscience, 1(1), 19–42. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18982117 [PubMed: 18982117]

Todorovic A, van Ede F, Maris E, & de Lange FP (2011, June). Prior Expectation Mediates Neural Adaptation to Repeated Sounds in the Auditory Cortex: An MEG Study. Journal of Neuroscience, 31(25), 9118–9123. doi: 10.1523/JNEUROSCI.1425-11.2011 [PubMed: 21697363]

Ungerleider LG, & Mishkin M. (1982, January). Two Cortical Visual Systems. In Ingle DJ, Goodale MA, & Mansfield RJW (Eds.), The Analysis of Visual Behavior (pp. 549–586). Cambridge, MA: MIT Press.

Urakubo H, Honda M, Froemke RC, & Kuroda S. (2008, March). Requirement of an allosteric kinetics of NMDA receptors for spike timing-dependent plasticity. The Journal of Neuroscience, 28(13), 3310–3323. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18367598 [PubMed: 18367598]

Usrey WM, & Sherman SM (2018). Corticofugal circuits: Communication lines from the cortex to the rest of the brain. Journal of Comparative Neurology, 0(0). doi: 10.1002/cne.24423

Valpola H. (2014, November). From neural PCA to deep unsupervised learning. arXiv:1411.7783 [cs, stat]. Retrieved 2017–05-15, from http://arxiv.org/abs/1411.7783

van Kerkoerle T, Self MW, Dagnino B, Gariel-Mathis M-A, Poort J, van der Togt C, & Roelfsema PR (2014, October). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. Proceedings of the National Academy of Sciences U.S.A, 111(40), 14332–14341. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25205811

VanRullen R. (2016, October). Perceptual cycles. Trends in Cognitive Sciences, 20(10), 723–735. doi: 10.1016/j.tics.2016.07.006 [PubMed: 27567317]

VanRullen R, & Koch C. (2003, May). Is perception discrete or continuous? Trends in Cognitive Sciences, 7(5), 207–213. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12757822 [PubMed: 12757822]

VanRullen R, & Thorpe SJ (2002, November). Surfing a spike wave down the ventral stream. Vision research, 42, 2593–2615. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12446033 [PubMed: 12446033]

Varela FJ, Toro A, John ER, & Schwartz EL (1981). Perceptual framing and cortical alpha rhythm. Neuropsychologia, 19(5), 675–686. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7312152 [PubMed: 7312152]

Vinken K, & Vogels R. (2017, November). Adaptation can explain evidence for encoding of probabilistic information in macaque inferior temporal cortex. Current Biology, 27(22), R1210–R1212. doi: 10.1016/j.cub.2017.09.018 [PubMed: 29161556]

von Stein A, Chiang C, & König P. (2000, December). Top-down processing mediated by interareal synchronization. Proceedings of the National Academy of Sciences of the United States of America, 97(26), 14748–14753. doi: 10.1073/pnas.97.26.14748 [PubMed: 11121074]

von Helmholtz H. (1867). Treatise on Physiological Optics, Vol III. Courier Corporation.

Waldert S, Lemon RN, & Kraskov A. (2013). Influence of spiking activity on cortical local field potentials. The Journal of Physiology, 591(21), 5291–5303. doi: 10.1113/jphysiol.2013.258228 [PubMed: 23981719]

Walsh KS, McGovern DP, Clark A, & O'Connell RG (2020, March). Evaluating the neurophysiological evidence for predictive processing as a model of perception. Annals of the New York Academy of Sciences, 1464(1), 242–268. doi: 10.1111/nyas.14321 [PubMed: 32147856]

Walter WG (1953). The living brain. Oxford, England: W. W. Norton.

Watanabe T, & Sasaki Y. (2015, January). Perceptual learning: Toward a comprehensive theory. Annual review of psychology, 66, 197–221. doi: 10.1146/annurev-psych-010814-015214

Whittington JCR, & Bogacz R. (2019, March). Theories of error back-propagation in the brain. Trends in Cognitive Sciences, 23(3), 235–250. doi: 10.1016/j.tics.2018.12.005 [PubMed: 30704969]

Williams RJ, & Zipser D. (1992, January). Gradient-based learning algorithms for recurrent networks and their computational complexity. In Chauvin Y. & Rumelhart DE (Eds.), Backpropagation: Theory, Architectures and Applications. Hillsdale, NJ: Erlbaum.

Wilson JR, Bose N, Sherman SM, & Guillery RW (1984, June). Fine structural morphology of identified X- and Y-cells in the cat's lateral geniculate nucleus. Proceedings of the Royal Society of London. Series B. Biological Sciences, 221(1225), 411–436. doi: 10.1098/rspb.1984.0042 [PubMed: 6146984]

Wimmer RD, Schmitt LI, Davidson TJ, Nakajima M, Deisseroth K, & Halassa MM (2015, October). Thalamic control of sensory selection in divided attention. Nature, 526(7575), 705–709. doi: 10.1038/nature15398 [PubMed: 26503050]

Wiskott L, & Sejnowski TJ (2002, April). Slow feature analysis: Unsupervised learning of invariances. Neural Computation, 14, 715–770. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11936959 [PubMed: 11936959]

Worden MS, Foxe JJ, Wang N, & Simpson GV (2000, March). Anticipatory biasing of visuospatial attention indexed by retinotopically specific alpha-band electroencephalography increases over occipital cortex. The Journal of neuroscience, 20. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10704517

Wurtz RH (2008, September). Neuronal mechanisms of visual stability. Vision Research, 48(20), 2070–2089. doi: 10.1016/j.visres.2008.03.021 [PubMed: 18513781]

Xing D, Yeh C-I, Burns S, & Shapley RM (2012, August). Laminar analysis of visually evoked activity in the primary visual cortex. Proceedings of the National Academy of Sciences, 109(34), 13871–13876. doi: 10.1073/pnas.1201478109

Yu C, & Smith LB (2012, November). Embodied attention and word learning by toddlers. Cognition, 125(2), 244–262. doi: 10.1016/j.cognition.2012.06.016 [PubMed: 22878116]

Zhou H, Schafer RJ, & Desimone R. (2016). Pulvinar-cortex interactions in vision and attention. Neuron, 89, 209–220. [PubMed: 26748092]

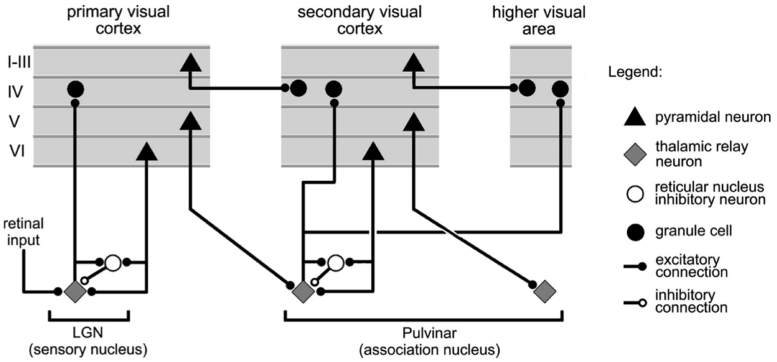

Figure 1:

Summary figure from Sherman & Guillery (2006) showing the strong feedforward *driver* projection emanating from layer 5IB cells in lower layers (e.g., V1), and the much more numerous feedback "modulatory" projection from layer 6CT (corticothalamic) cells. We interpret these same connections as providing a prediction (6CT) vs. outcome (5IB) activity pattern over the pulvinar.

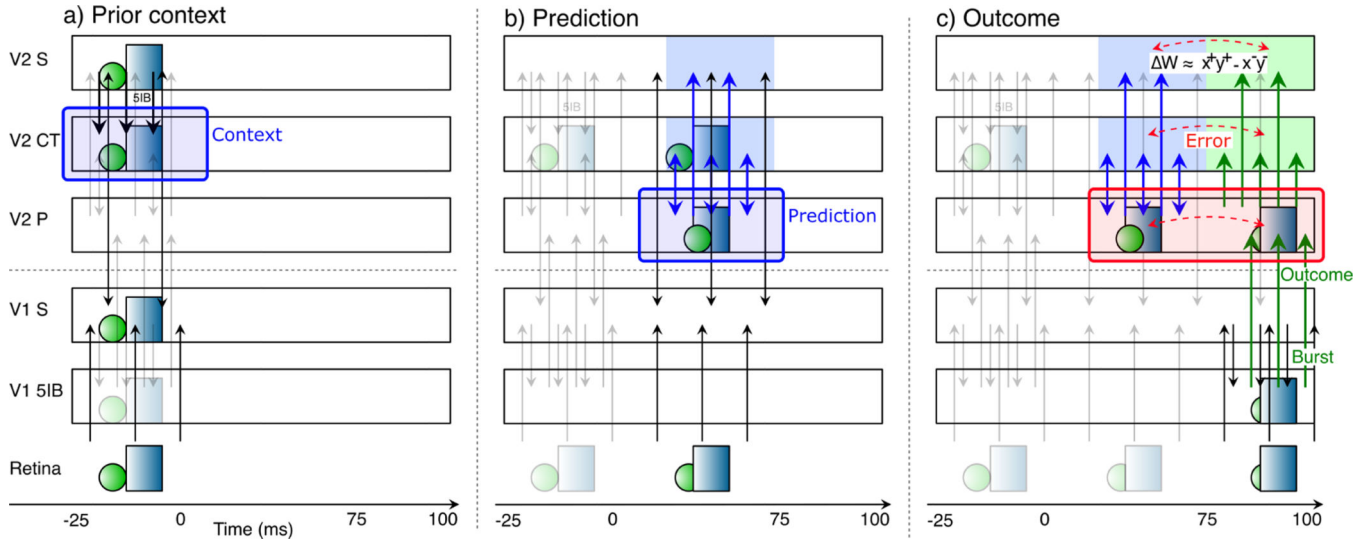

Figure 2:

Corticothalamic information flow under our predictive learning hypothesis, shown as a sequence of movie frames (Retina), illustrating the three key steps taking place within a single 125 ms time window, broken out separately across the three panels: **a)** prior context is updated in the V2 CT layer; **b)** which is then used to generate a prediction over the pulvinar (V2 P); **c)** against which the outcome, driven by bottom-up 5IB bursting, represents the prediction error *as a temporal difference between the prediction and outcome states over the pulvinar*. Changes in synaptic weights (learning) in all superficial (S) and CT layers are driven from the *local* temporal difference experienced by each neuron, using a form of the contrastive hebbian learning (CHL) term as shown, where the '+' superscripts indicate outcome activations, and '−' superscripts indicate prediction. CHL approximates the backpropagated prediction error gradient experienced by each neuron (O'Reilly, 1996), reflecting both direct pulvinar error signals, and indirect corticocortical error signals as well. In specific: **a)** CT context updating occurs via 5IB bursting (not shown) in higher layer (V2) during prior alpha (100 ms) cycle — this context is maintained in the CT layer and used to generate predictions. **b)** The prediction over pulvinar is generated via numerous top-down CT projections. This prediction state also projects up to S and CT layers, and from S to all other S layers via extensive bidirectional connectivity, so their activation state reflects this prediction as well. **c)** The subsequent outcome drives pulvinar activity bottom-up via V1 5IB bursting, and is likewise projected to S and CT layers, ensuring that the relevant temporal difference error signal is available locally in cortex. The difference in activation values across these two time points, in S and CT layers throughout the network, drives learning to reduce prediction errors. Note that the single most important property of the 5IB bursting is that these driver cells are *not* active during the prediction phase — the bursting itself may also be useful in the driving property, but that is a secondary consideration to the critical feature of having a time when the prediction alone can be projected onto the pulvinar.

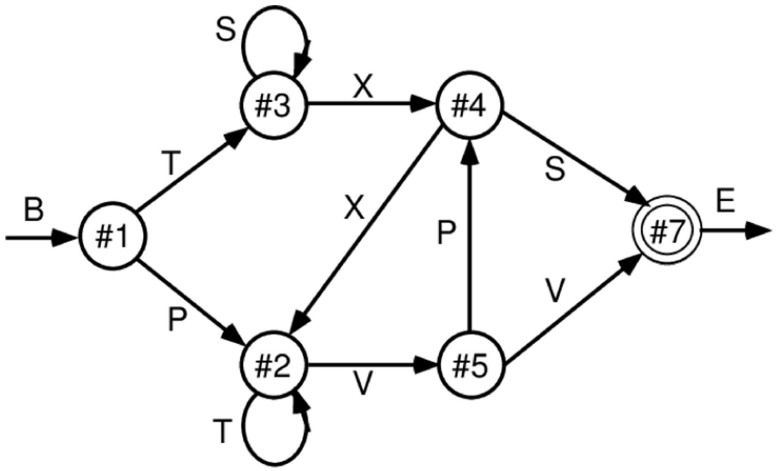

Figure 3:

Finite state automaton (FSA) grammar used in implicit sequential learning experiments (Reber, 1967) and in early simple recurrent networks (SRNs) (Cleeremans & McClelland, 1991). It generates a sequence of letters according to the link transitioned between state nodes, where each outgoing link to another node has a 50% probability of being selected. Each letter (except for the B=begin and E=end) appears at 2 different points in the grammar, making them locally ambiguous. This combination of randomness and ambiguity makes it challenging for a learning system to infer the true underlying structure of the grammar.

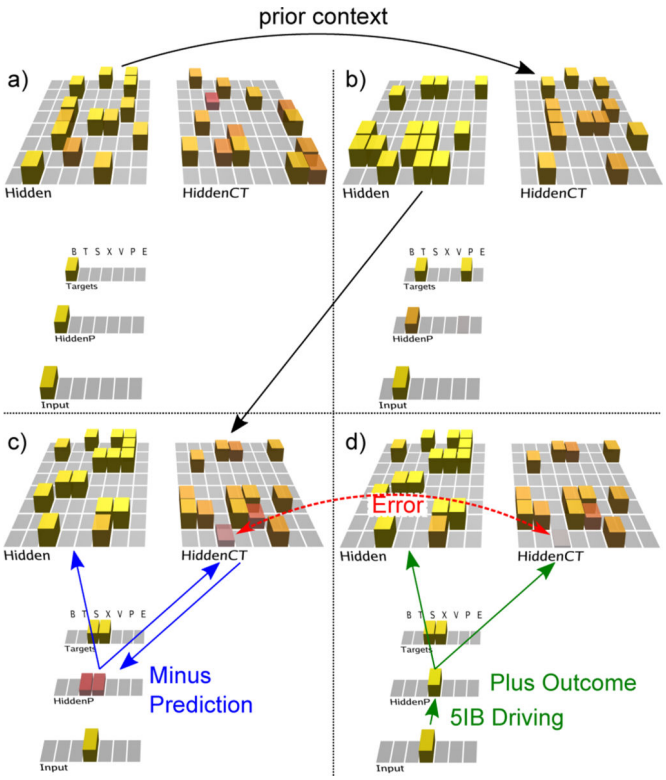

Figure 4:

Predictive learning model applied to the FSA grammar shown in previous figure. The first three panels (a, b, c) show the prediction state (end of the *minus* phase, e.g., the first 75 ms of an alpha cycle) of the trained model on the first three steps of the sequence 'BTX' (plus phases also occurred, but are not shown). The last panel (d) shows the plus phase after the third step. The *Input* layer provides the 5IB drivers for the corresponding *HiddenP* pulvinar layer, so the plus phase is always based on the specific randomly-selected path taken. The *Targets* layer is purely for display, showing the two valid possible labels that could have been predicted. To track learning, the model's prediction is scored as accurate if either or both targets are activated. Computationally, the model is similar to an SRN, where the CT layer that drives the prediction over the pulvinar encodes the activation state from the previous time step (alpha cycle), due to the phasic bursting of the 5IB neurons that drive CT updating. Note how the CT layer in b) reflects the Hidden activation state in a), and likewise for c) reflecting b). This is evident because we're using one-to-one connectivity between Hidden and HiddenCT layers (which works well in general, along with full lateral connectivity within the CT layer). Thus, even though the correct answer is always present on the Input layer for each step, the CT layer is nevertheless attempting to predict this Input based on the information from the prior time step. **a)** In the first step, the B label is unambiguous and easily predicted (based on prior E context). **b)** In the 2nd step, the network correctly guesses that the T label will come next, but there is a faint activation of the other P alternative, which is also activated sometimes based on prior learning history and associated minor weight tweaks. **c)** In the 3rd step, both S and X are equally predicted. **d)** In the *plus* phase, only the Input pattern ('X' on this trial) drives HiddenP activations, and the projections from pulvinar back to the cortex convey both the minus-phase prediction and

plus-phase actual input. You can see one HiddenCT neuron, just above the arrow, visibly changes its activation as a result (and all neurons experience smaller changes), and learning in all these cortical (Hidden) layer neurons is a function of their local temporal difference between minus and plus phases.

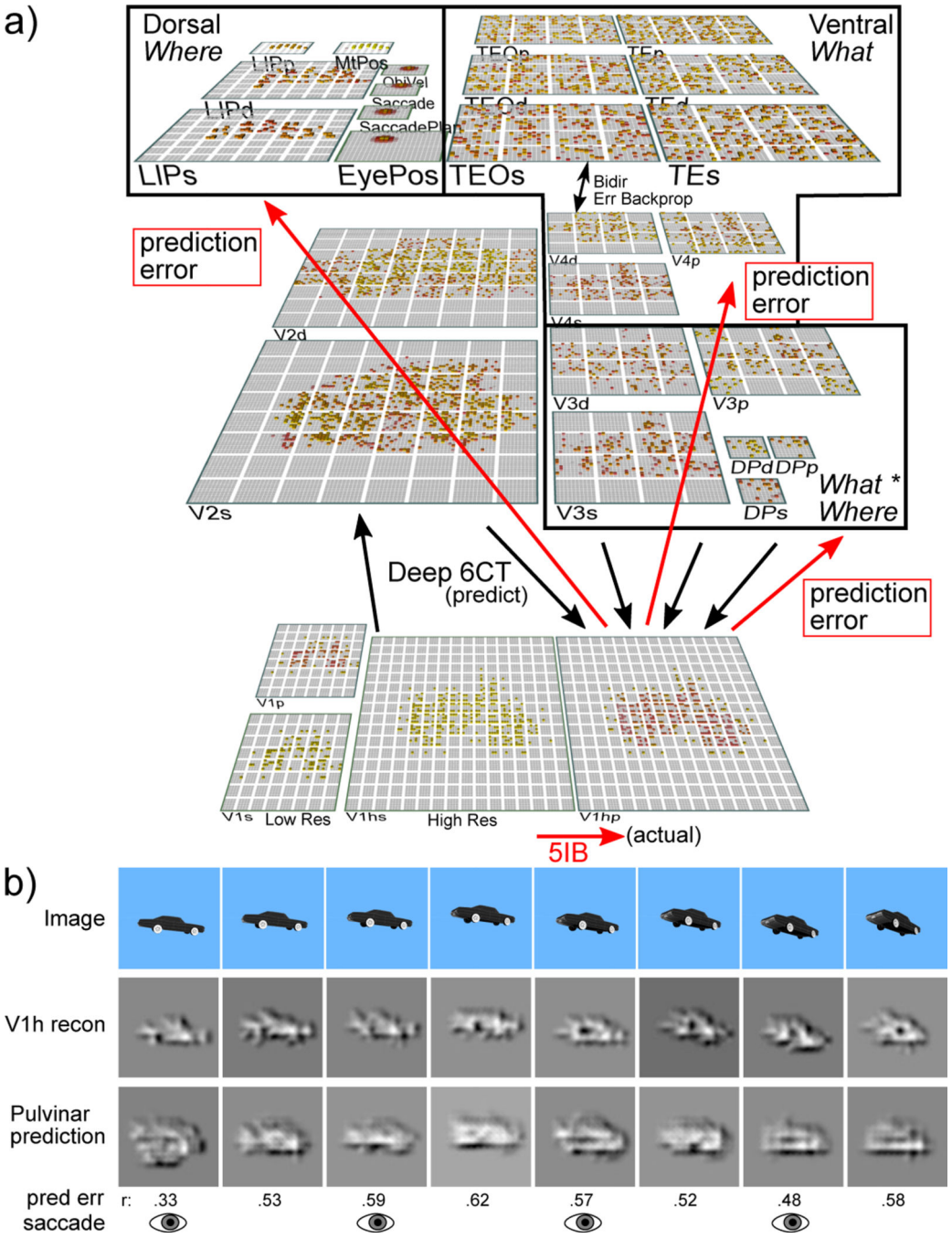

Figure 5:

a) The *What-Where-Integration, WWI* deep predictive learning model. The dorsal *Where* pathway learns first, using easily-abstracted *spatial blobs*, to predict object location based on prior motion, visual motion, and saccade efferent copy signals. This drives strong top-down inputs to lower areas with accurate spatial predictions, leaving the *residual* error concentrated on *What* and *What * Where* integration. The V3 and DP (dorsal prelunate) constitute the *What * Where* integration pathway, binding features and locations. V4, TEO, and TE are the *What* pathway, learning abstracted object category representations, which

also drive strong top-down inputs to lower areas. Suffixes: *s* = superficial, *d* = deep, *p* = pulvinar. **c)** Example sequence of 8 alpha cycles that the model learned to predict, with the reconstruction of each image based on the V1 gabor filters (*V1h recon*), and model-generated prediction (correlation *r* prediction error shown). The low resolution and reconstruction distortion impair visual assessment, but *r* values are well above the *r*'s for each V1 state compared to the previous time step (mean = .38, min of .16 on frame 4 — see Appendix for more analysis). Eye icons indicate when a saccade occurred.

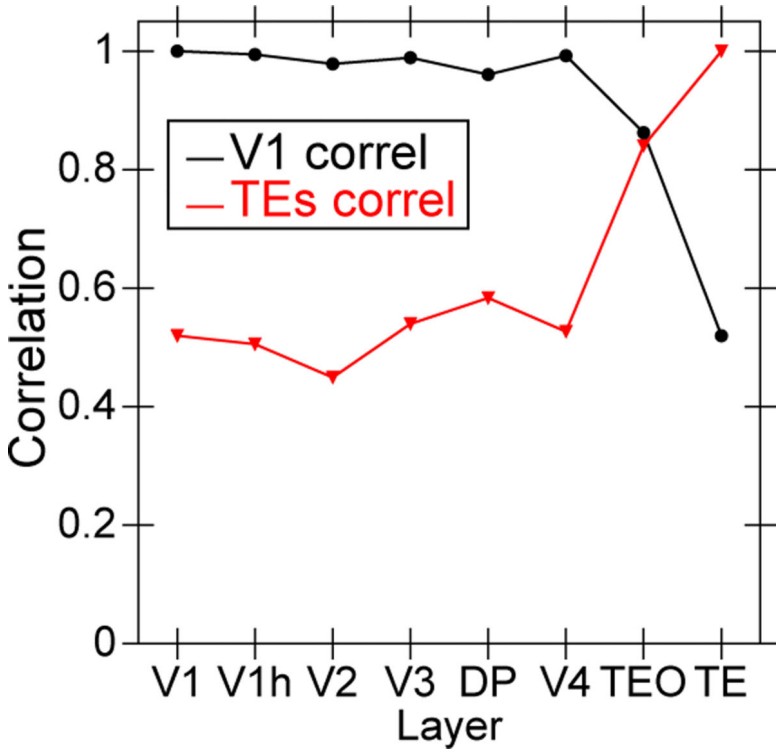

Figure 6:

Emergence of abstract category structure over the hierarchy of layers, comparing similarity structure in each layer vs that present in V1 (black line) or in TE (red line). Both cases, which are roughly symmetric, clearly show that IT layers (TEO, TE) progressively differentiate from raw input similarity structure present in V1, and, critically, that the model has learned structure beyond that present in the input. This is is the simplest, most objective summary statistic showing this progressive emergence of structure, while subsequent figures provide a more concrete sense of what kinds of representations actually developed.

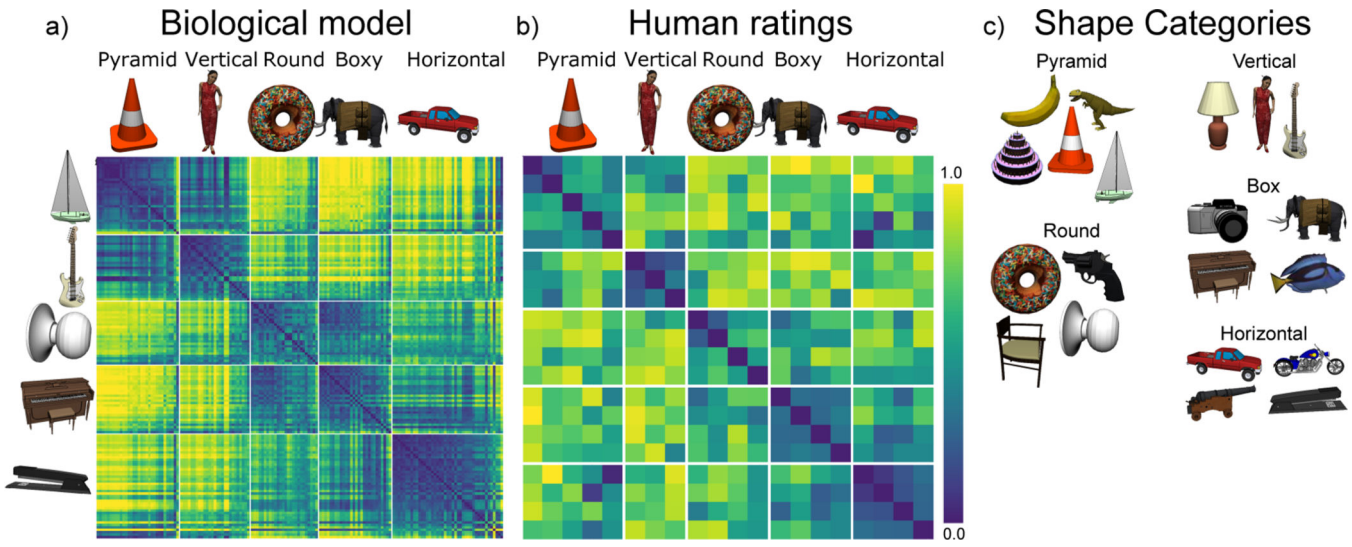

Figure 7:

a) Category similarity structure that emerged in the highest layer, TE, of the biologically based predictive learning model, showing *dissimilarity* (1-correlation) of the TE representation for each 3D object against every other 3D object (156 total objects). Blue cells have high similarity. Model has learned block-diagonal clusters or categories of high-similarity groupings, contrasted against dissimilar off-diagonal other categories. Clustering maximized average *within – between* dissimilarity (see Appendix), and clearly corresponded to the shown shape-based categories, with exemplars from each category shown. Also, all items from the same basic-level object categories (N=20) are reliably subsumed within learned categories. **b)** Human similarity ratings for the same 3D objects, presented with the V1 reconstruction (see Fig 1b) to capture coarse perception in the model, aggregated by 20 basic-level categories (156 × 156 matrix was too large to sample densely experimentally). Each cell is 1 proportion of time given object pair was rated more similar than another pair (see Appendix). The human matrix shares the same centroid categorical structure as the model (confirmed by permutation testing and agglomorative cluster analysis, see Appendix), indicating that human raters used the same shape-based category structure. **c)** One object from each of the 20 basic level categories, organized into the shape-based categories. The Vertical, Box and Horizontal categories are fairly self-evident and the model was most consistent in distinguishing those, along with subsets of the Pyramid (layer-cake, traffic-cone, sailbot) and Round (donut, doorknob) categories, while banana, trex, chair, and handgun were more variable.

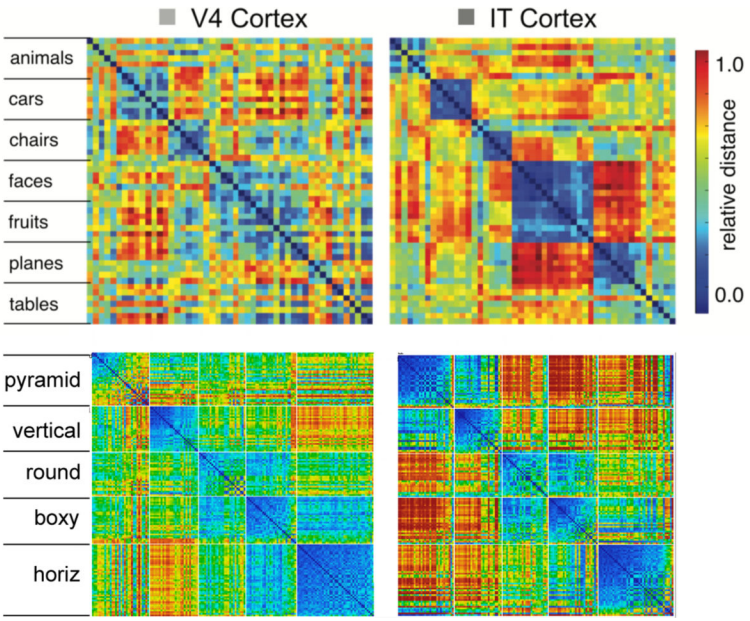

Figure 8:

Comparison of progression from V4 to IT in macaque monkey visual cortex (top row, from Cadieu et al., 2014) versus same progression in model (replotted using comparable color scale). Although the underlying categories are different, and the monkeys have a much richer multi-modal experience of the world to reinforce categories such as foods and faces, the model nevertheless shows a similar qualitative progression of stronger categorical structure in IT, where the block-diagonal highly similar representations are more consistent across categories, and the off-diagonal differences are stronger and more consistent as well (i.e., categories are also more clearly differentiated). Note that the critical difference in our model versus those compared in Cadieu et al. 2014 and related papers is that they explicitly trained their models on category labels, whereas our model is *entirely self-organizing* and has no external categorical training signal.

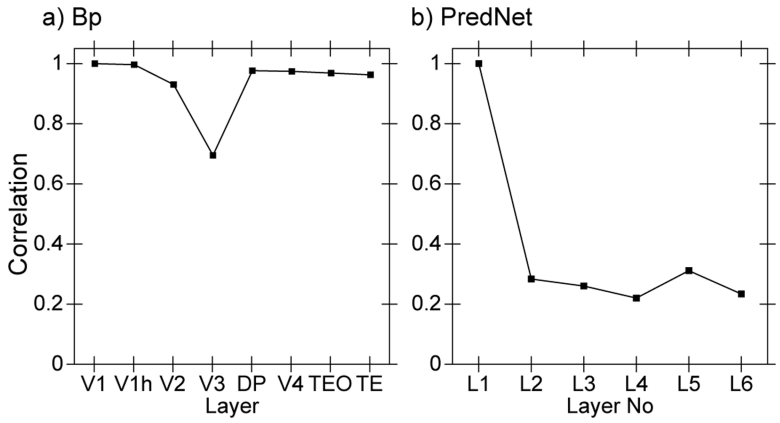

Figure 9:

Similarity of similarity structure across layers for the comparison backprop models, comparing each layer to the first layer. **a)** Backpropagation (Bp) model with the same What / Where structure as the biological model. Unlike the biologically based model (Figure 6) the higher IT layers (TE, TEO) do not diverge significantly from the similarity structure present in V1, indicating that the model has not developed abstractions beyond the structure present in the visual input. Layer V3 is most directly influenced by spatial prediction errors, so it differs from both in strongly encoding position information. **b)** PredNet model, which has 6 layers. Layers 2–6 diverge from layer 1, but there is no progressive change in the higher layers as we see in our model moving from V4 to TEO. The divergence in correlation starting at layer 2 is likely due to the fact that higher layers only encode errors, not stimulus-driven positive representations of the input. Aside from this large distinction (which is inconsistent with the similarity in neural coding seen in actual V1 and V2 recordings), there is no evidence of a cumulative development of abstraction in higher layers.

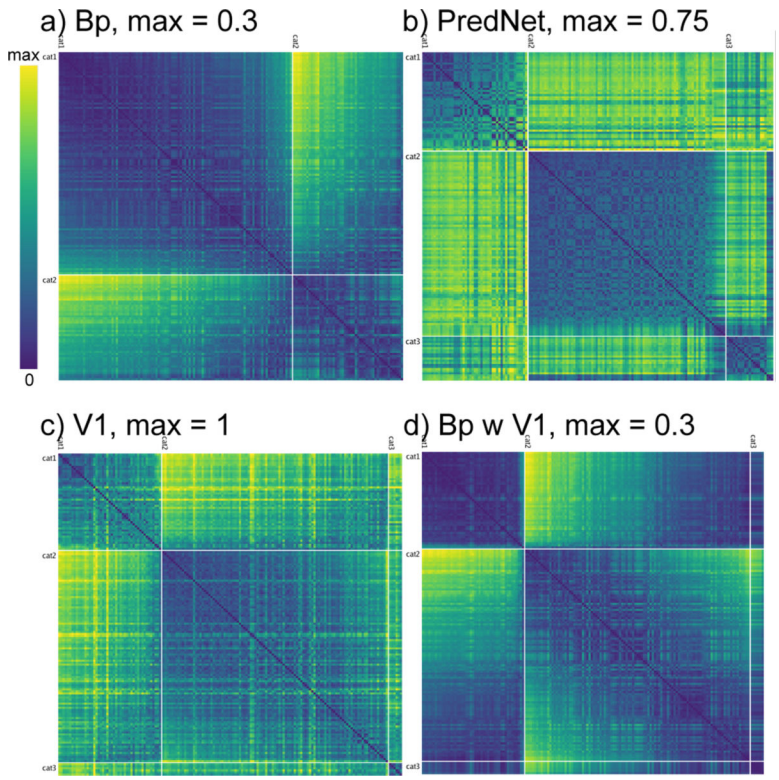

Figure 10:

a) Best-fitting category similarity for TE layer of the backpropagation (Bp) model with the same What / Where structure as the biological model. Only two broad categories are evident, and the lower *max* distance (0.3 vs. 1.5 in biological model) means that the patterns are much less differentiated overall. **b)** Best-fitting similarity structure for the PredNet model, in the highest of its layers (layer 6), which is more differentiated than Bp (max = 0.75) but also less cleanly similar within categories (i.e., less solidly blue along the block diagonal), and overall follows a broad category structure similar to V1. **c)** The best fitting V1 structure, which has 2 broad categories and banana is in a third category by itself. The lack of dark blue on the block diagonal indicates that these categories are relatively weak, and every item is fairly dissimilar from every other. **d)** The Bp TE similarity values from panel a shown in the same ordering as V1 from panel c, demonstrating how the similarity structure has not diverged very much, consistent with the results shown in Figure 9 — the within – between contrast differences are 0.0838 for panel a and 0.0513 for d — see Appendix for details.

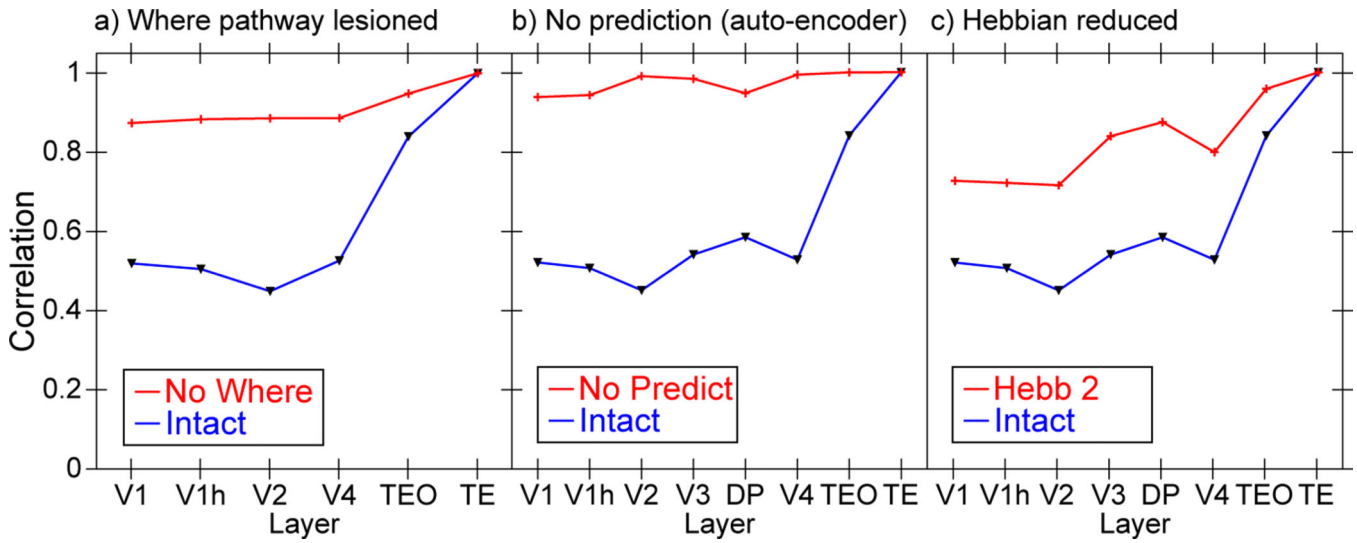

Figure 11:

Effects of various manipulations on the extent to which TE representations differentiate from V1. For all plots, *Intact* is the same result shown in Figure 6 from the intact model for ease of comparison (panel a is missing V3 and DP dorsal pathway layers). All of the following manipulations significantly impair the development of abstract TE categorical representations (i.e., TE is more similar to V1 and the other layers). **a)** Dorsal *Where* pathway lesions, including lateral inferior parietal sulcus (LIP), V3, and dorsal prelunate (DP). This pathway is essential for regressing out location-based prediction errors, so that the residual errors concentrate feature-encoding errors that train the *What* pathway. **b)** Allowing the deep layers full access to current-time information, thus effectively eliminating the prediction demand and turning the network into an auto-encoder, which significantly impairs representation development, and supports the importance of the challenge of predictive learning for developing deeper, more abstract representations. **c)** Reducing the strength of Hebbian learning by 20% (from 2.5 to 2), demonstrating the essential role played by this form of learning on shaping categorical representations. Eliminating Hebbian learning entirely (not shown) prevented the model from learning anything at all, as it also plays a critical regularization and shaping role on learning.

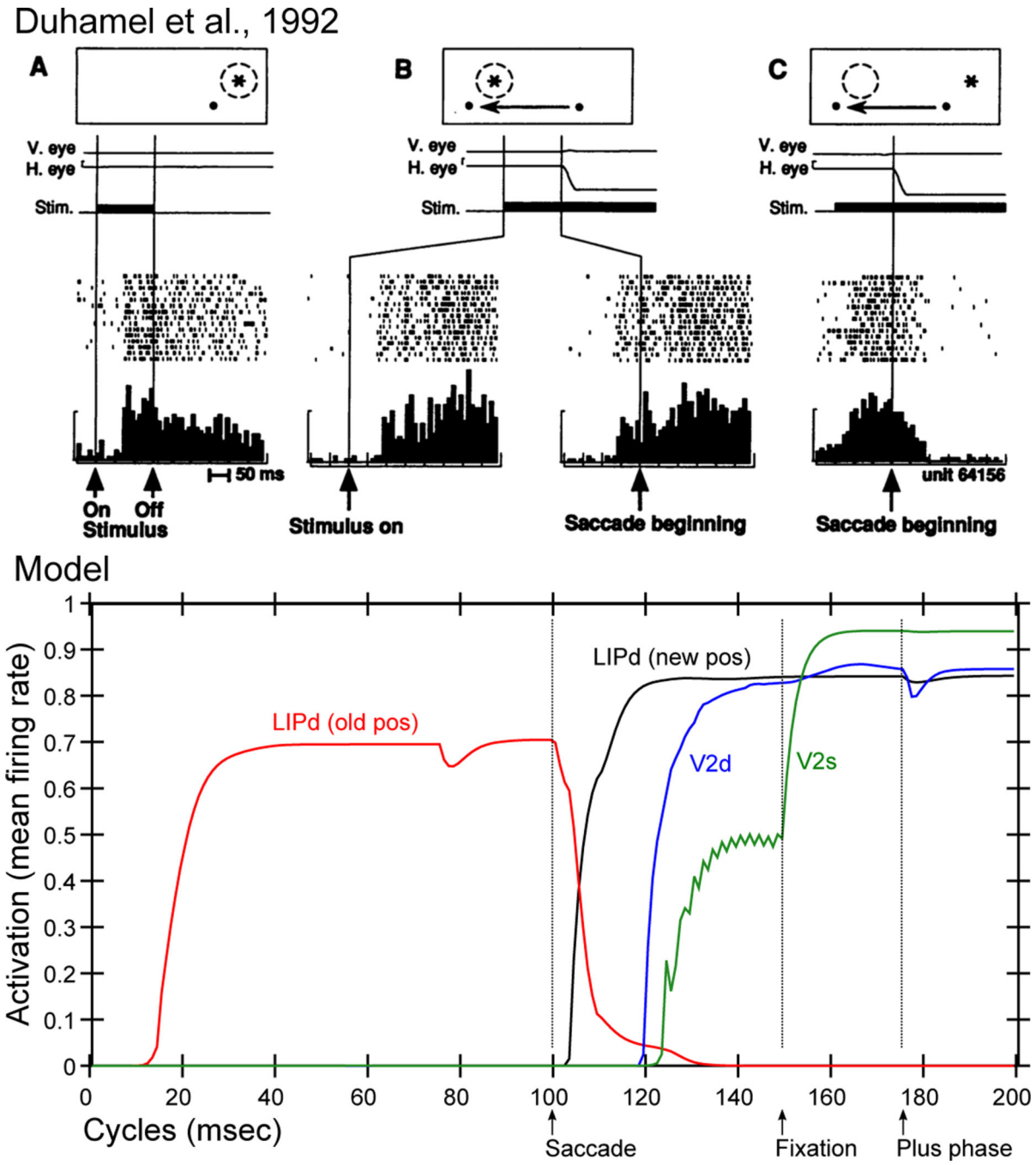

Figure 12:

Predictive Remapping. **top:** Original remapping data in LIP from Duhamel et al (1992). A) shows stimulus (star) response within receptive field (dashed circle) relative to fixation dot (upper right of fixation). B) Just prior to monkey making a saccade to new fixation (moving left), stimulus is turned on in receptive field location that *will be* upper right of the new fixation point, and the LIP neuron responds to that stimulus in advance of the saccade completing. The neuron does not respond to the stimulus in that location if it is not about to make a saccade that puts it within its receptive field (not shown). This is

predictive remapping. C) response to the old stimulus location goes away as saccade is initiated. **bottom:** Data from our model, from individual units in LIPd, V2d, and V2s, showing that the LIP deep neurons respond to the saccade first, activating in the new location and deactivating in the old, and this LIP activation goes top-down to V3 and V2 to drive updating there, generally at a longer latency and with less activation especially in the superficial layers. When the new stimulus appears at the point of fixation (after a 50 ms saccade here), the *primed* V2s units get fully activated by the incoming stimulus. But the deep neurons are insulated from this superficial input until the plus phase, when the cascade of 5IB firing drives activation of the actual stimulus location into the pulvinar, which then reflects up into all the other layers.

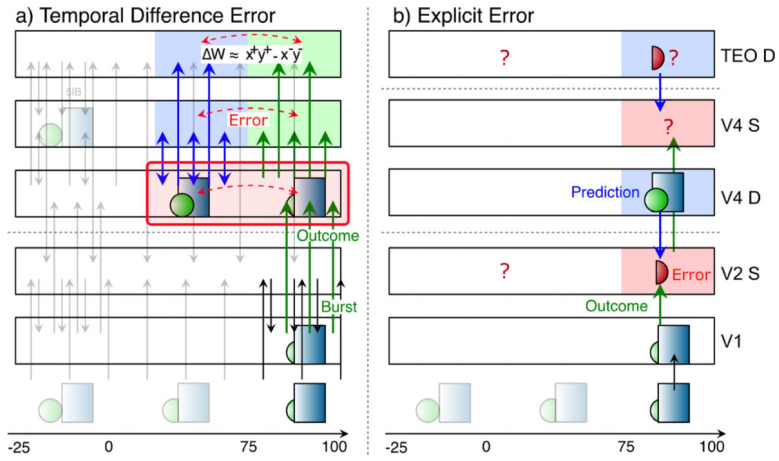

Figure 13:

Comparison between: **a)** The proposed thalamocortical temporal-difference predictive learning model (from Figure 2), versus **b)** The Bayesian-style explicit error (EE) coding model (Rao & Ballard, 1999; Friston, 2010, Bastos et al., 2012). The EE model holds that superficial (S, lamina 2/3) error-coding neurons receive the prediction via a net inhibitory top-down projection from higher-level deep layer (D) neurons, and an excitatory bottom-up projection representing the outcome, such that their activation represents the difference. To encode both signs of the error (omissions, false alarms) with positive-only spike rates, two separate populations of EE neurons would be required, or a more complicated deviation from tonic firing level scheme. Unambiguous evidence of such EE coding neurons has not been found (Walsh et al, 2020). In contrast, error signals in our proposed framework remain as a temporal difference between the two states of prediction vs. outcome, *which enables all connectivity between cortical areas to be excitatory and always represent a positive encoding of either the prediction or outcome*. In contrast, under EE, after one error subtraction at the lowest level, only error signals are hypothesized to flow forward to higher layers, meaning that the representations at higher layers are about increasingly higher-order *errors*, not positive encodings of the environmental state at increasing levels of abstraction. These are indicated by ? because they are difficult to picture intuitively, and they are inconsistent with extensive available data showing similar positive representations of the external world at all levels in the visual hierarchy. Although some frameworks make claims about temporal dynamics, these are not strongly constrained by the basic computational framework, so that also remains a question.