Deep Predictive Learning in Neocortex and Pulvinar

Randall C. O'Reilly, Jacob L. Russin, and John Rohrlich
Department of Psychology, Computer Science, and Center for Neuroscience
University of California Davis
1544 Newton Ct
Davis, CA 95618
oreilly@ucdavis.edu

May 30, 2020

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.

Abstract

How does the human brain learn new concepts from raw sensory experience, without explicit instruction? We still do not have a widely-accepted answer to this central question. Here, we propose a detailed biological mechanism for the widely-embraced idea that learning is based on 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, strong *driver* inputs from lower areas supply the actual outcome, originating in layer 5 intrinsic bursting (5IB) neurons. Thus, the outcome is only briefly activated, roughly every 100 msec (i.e., 10 Hz, *alpha*), resulting in a *temporal difference error signal*, which drives local synaptic changes throughout the neocortex, resulting 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. There have been significant advances in understanding the detailed biochemical basis of learning in terms of synaptic plasticity between neurons (Lüscher & Malenka, 2012), and many cognitive and computational models of learning. However, there is still no widely-accepted answer to this puzzle, that is clearly supported by known biological mechanisms and also produces effective learning at computational and cognitive levels. At these functional levels, the idea that we learn via an active *predictive* process goes back to Helmholtz's *recognition by synthesis* proposal (von Helmholtz, 2013), and has been widely embraced in a wide range of different frameworks (Elman, 1990; Elman, Bates, Karmiloff-Smith, Johnson, Parisi, & Plunkett, 1996; Mumford, 1992; Dayan, Hinton, Neal, & Zemel, 1995; Rao & Ballard, 1999; Kawato, Hayakawa, & Inui, 1993; Friston, 2005).

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 pulvinar nucleus of the thalamus (Sherman & Guillery, 2006; Usrey & Sherman, 2018). Specifically, numerous weak projections into the thalamic relay cells (TRCs) in the pulvinar drive top-down predictions, and sparse, strong *driver* inputs from lower areas encode the actual outcome, and learning is based on the difference. Because these driver inputs originate in layer 5 intrinsic bursting (5IB) neurons, the outcome is only briefly activated, roughly every 100 msec (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 learning (O'Reilly, 1996; Ackley, Hinton, & Sejnowski, 1985; Hinton & McClelland, 1988; Bengio, Mesnard, Fischer, Zhang, & Wu, 2017; Whittington & Bogacz, 2019; Lillicrap, Santoro, Marris, Akerman, & Hinton, 2020).

One primary objective of this paper is to describe this biologically-based mechanism for predictive error-driven learning in sufficient detail that it can be clearly evaluated 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, a major contribution of this work is to provide a clear functional role for this distinctive thalamocortical circuitry that contrasts with existing ideas about what it might be doing, in testable ways.

A second major objective is to implement this predictive error-driven learning mechanism in a computational model that faithfully captures its essential biological features, while still being sufficiently simplified computationally that it can be used to simulate large-scale brain networks, to test whether the learning mechanism can drive the formation of cognitively-useful representations. In particular, there is a critical question for any purely predictive-learning model: can it develop high-level, abstract ways of representing the raw sensory inputs, while learning from nothing but predicting these low-level visual inputs. For example, most current models of visual object recognition that have been compared against neurophysiological data rely on large human-labeled image datasets to explicitly train abstract category information via error-backpropagation (Cadieu, Hong, Yamins, Pinto, Ardila, Solomon, Majaj, & DiCarlo, 2014; Rajalingham, Issa, Bashivan, Kar, Schmidt, & DiCarlo, 2018). Existing predictive-learning models based on error backpropagation (Lotter, Kreiman, & Cox, 2016) have not clearly demonstrated the development of abstract, categorical representations without additional human-labeled training. Instead, previous work has shown that predictive learning can be a useful method for pretraining networks that are subsequently trained using human-generated labels.

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 systematically categorize 3D objects according to invariant shape properties, based on raw visual inputs alone. We found that these categories match human judgments on the same stimuli, and are consistent with neural representations in inferotemporal (IT) cortex in primates (Cadieu et al., 2014). Furthermore, we show that

comparison predictive DCNN models lacking these biological features (Lotter et al., 2016) did not learn object categories that go beyond the visual input structure. Thus, it is possible that incorporating certain biological properties of the brain can potentially provide a better understanding of human learning at multiple levels relative to existing DCNN models. However, it is important to emphasize that our objectives in this work are *not* to produce a better machine-learning (ML) algorithm per se, but rather to test the computational properties of our biologically-based, scientific theory for how the mammalian brain might learn. 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: it should instead 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. Next, we discuss the relevant biological data in detail, along with testable predictions that can differentiate this account of what this system does relative to existing ideas. Then, we present the large-scale model of the visual system, which learns by predicting over brief visual movies of 3D objects rotating and translating over time and space. We find that the model develops strongly categorical, shape-based representations in its upper IT layers, and these match those of human participants evaluating the same 3D objects. Furthermore, we show that these categorical representations diverge significantly from the similarity structure present in the lower layers of the network. Thus, we conclude that this form of predictive error-driven learning is capable of going beyond the surface structure of the raw sensory input, to develop higher-level abstract representations that otherwise have only been produced in neural models through explicit training via human-labeled image datasets. To further explore this space, we evaluated two other prediction-error learning models using pure error-backpropagation, based on current deep-convolutional neural network (DCNN) principles, and found that they did not develop the same kind of high-level categories, and instead remained largely tied to the similarity structure of the raw visual inputs. 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.

Predictive Error-driven Learning in the Neocortex and Pulvinar

Figure 1a shows the thalamocortical circuits characterized by Sherman and Guillery (2006) (see also 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; (Shipp, 2003; Arcaro, Pinsk, & Kastner, 2015). One projection consists of numerous, weaker connections originating in deep layer VI of the neocortex (the 6CT corticothalamic projecting cells). The other is a very sparse (typically one-to-one; Rockland, 1998, 1996) and very strong *driver* pathway that originates from lower-level layer 5 intrinsic bursting cells (5IB). These 5IB neurons fire discrete bursts roughly every 100 msec (Larkum, Zhu, & Sakmann, 1999; Franceschetti, Guatteo, Panzica, Sancini, Wanke, & Avanzini, 1995; Lorincz, Kekesi, Juhasz, Crunelli, & Hughes, 2009; Saalmann, Pinsk, Wang, Li, & Kastner, 2012), which corresponds to 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; VanRullen & Koch, 2003; Jensen, Bonnefond, & VanRullen, 2012; Fiebelkorn & Kastner, 2019).

The existing literature generally characterizes the 6CT projection as *modulatory* (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 5IB inputs (Bender, 1982; Petersen, Robinson, & Keys, 1985; Bender & Youakim, 2001; Robinson, 1993; Saalmann et al., 2012; Komura, Nikkuni, Hirashima, Uetake, & Miyamoto, 2013; Zhou, Schafer, & Desimone, 2016). Indeed, these recordings show

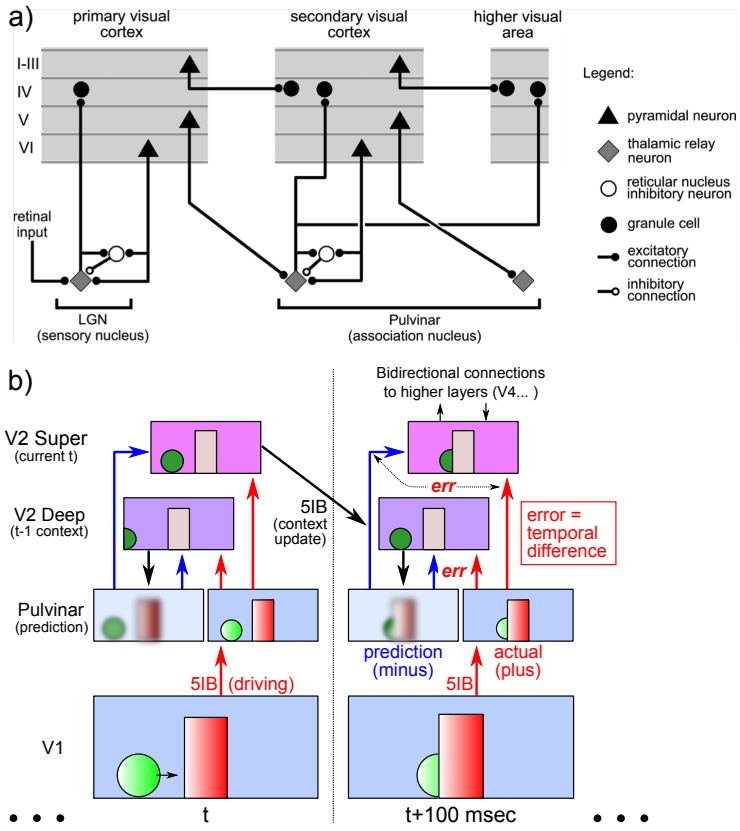

Figure 1: **a)** Summary figure from Sherman & Guillery (2006) showing the strong feedforward driver projection emanating from layer 5IB cells in lower layers (e.g., V1), and the much more numerous feedback "modulatory" projection from layer 6CT cells. We interpret these same connections as providing a prediction (6CT) vs. outcome (5IB) activity pattern over the pulvinar. **b)** Temporal evolution of information flow under our prediction error hypothesis, operating on visual sequences, over two alpha cycles of 100 msec each. In each alpha cycle, the V2 Deep layer (lamina 5, 6) uses the prior 100 msec of context to generate a prediction (*minus* phase) on the pulvinar thalamic relay cells (TRC). The bottom-up outcome is driven by V1 5IB strong driver inputs (*plus* phase); error-driven learning occurs as a function of the *temporal difference* between these phases, in both superficial (lamina 2, 3) and deep layers, sent via broad pulvinar projections. 5IB bursting in V2 drives update of temporal context in V2 Deep layers, and also the plus phase in higher area TRC, to drive higher-level predictive learning.

that pulvinar neural firing generally resembles that of the visual areas they interconnect with. This is important, because our predictive learning framework requires that these 6CT top-down projections be capable of driving TRC activity directly.

Specifically, in contrast to the standard view, the core idea behind our theory is that the top-down 6CT projections drive a *prediction* across the extent of the pulvinar, which precedes the subsequent *outcome* state resulting from the strong 5IB driver inputs, as illustrated in Figure 1b. This prediction activity state in pulvinar TRC neurons should develop during the first roughly 75 msec of a 100 msec alpha cycle, while the final 25 msec largely reflects the strong 5IB bottom-up ground-truth driver inputs. Thus, the difference or prediction error signal is reflected in the temporal difference of these activation states over time. In other words, our hypothesis is that the pulvinar is only ever directly representing either the top-down prediction or the bottom-up actual outcome, and the prediction-error difference between these remains as an implicit

difference in these activation states over time. Note that in the typical lab experiment with phasic stimuli presented outside of an ongoing predictable temporal sequence (which is uncharacteristic of the natural world), there may often be no significant prediction prior to stimulus onset, in which case a 5IB burst is observed immediately on stimulus onset. More generally, strong transient stimuli likely drive 5IB bursting to reset the phase of the alpha cycle, as we discuss in greater detail later.

The properties of these two pathways are notably well suited for this predictive learning role, in the following ways:

- A true prediction (i.e., about the future, as in the famous quote about what makes prediction hard; "prediction is very difficult, especially about the future", attributable to Danish author Robert Storm Petersen) must be prevented from cheating and relying on direct information about that which is being predicted: thus there must be a mechanism preventing the outcome information from "contaminating" the prediction. The phasic, bursting nature of the 5IB driver inputs provides this essential feature, giving a window where no outcome signals are present, when the prediction can be represented.

- Generating a prediction requires converging inputs from a range of higher-level cortical areas, to integrate the contributions of multiple different specialized pathways in the challenging problem of predicting what will happen next. This is consistent with the broad, integrative nature of the top-down 6CT inputs (Shipp, 2003; Mumford, 1991).

- Furthermore, it can take some time to integrate all these signals, which is consistent with the outcome bursts occupying a briefer 25 msec of the 100 msec alpha cycle, with the remainder available for this integration. The overall duration of the alpha cycle itself may represent a reasonable compromise between this integration time and the need to keep up with tracking changes in the world — 100 msec is otherwise a relatively slow time period relative to the synaptic integration time constants in the basic thalamocortical circuit itself.

- The outcome signal should be as *veridical* as possible, and should arise from lower areas in the hierarchy relative to the corresponding 6CT inputs: the bottom-up, one-to-one nature of the 5IB driver projections can directly convey such veridical outcome signals.

- The prediction error signal should be widely broadcast back out to the same areas that provide the top-down predictions, to provide the training signal that improves these predictions. This is also a known, distinctive property of this circuitry (Shipp, 2003; Mumford, 1991).

- For cortical neurons receiving these projections from the pulvinar, there must be some way in which the difference between prediction and outcome (i.e., the error itself) can drive learning. Here we hypothesize that this difference remains as a *temporal difference* error signal, i.e., the difference over time in pulvinar activation states, arising naturally as a prediction state followed by the outcome state. This 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 (Rao & Ballard, 1999; Kawato et al., 1993; Friston, 2005, 2010; Ouden, Kok, & Lange, 2012; Lotter et al., 2016). Despite many attempts to identify such explicit error-coding neurons in the cortex, no strong evidence has been found (Kok & de Lange, 2015; Kok, Jehee, & de Lange, 2012; Summerfield & Egner, 2009; Lee & Mumford, 2003; 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. Thus, we think that the temporal-difference nature of the prediction error signal is more efficient and should naturally emerge from the basic operation of the circuit.

- Furthermore, there is a long history of computational models of error-driven learning based on temporal-difference signals (Ackley et al., 1985; O'Reilly, 1996), and we have recently provided a direct biological mechanism for this form of learning based on a biologically-detailed model of spike timing dependent plasticity (STDP) (Urakubo, Honda, Froemke, & Kuroda, 2008). We showed that when activated by realistic Poisson spike trains, this STDP model produces a non-monotonic learning curve similar to that of the BCM model (Bienenstock, Cooper, & Munro, 1982), which has been widely established as resulting from competing calcium-driven postsynaptic plasticity pathways (Lüscher & Malenka, 2012). 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, McKee, Woloszyn, Amit, Freedman, Sheinberg, & Brunel, 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, which mathematically approximates gradient descent to minimize overall prediction errors (O'Reilly, 1996).

Thus, remarkably, this thalamocortical circuit appears to provide *precisely* the necessary ingredients to support predictive error-driven learning. Interestingly, although (Sherman & 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 (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 explict error-coding framework and not involving any aspects of the above-described circuitry.

As we discuss later, this proposed predictive role for the pulvinar is not incompatible with the more widely-discussed role it may play in attention (LaBerge & Buchsbaum, 1990; Bender & Youakim, 2001; Snow, Allen, Rafal, & Humphreys, 2009; Saalmann & Kastner, 2011; Zhou et al., 2016; Fiebelkorn & Kastner, 2019). Indeed, we think these two functions are synergistic (i.e., you predict what you attend, and vice-versa), and have initial computational results consistent with this idea.

In the following section, we discuss some of the most important neural data of relevance to our hypotheses (beyond that summarized above) followed by a list of some predictions that would clearly test the validity of this framework.

Existing Neuroscience Data

Extensive biological evidence supports the alpha-frequency dynamics of the deep layer network, in contrast to a dominant gamma frequency for the superficial layers, corresponding to the 25 msec subdivision of the overall alpha cycle. This includes direct electrophysiological recording (Luczak, Bartho, & Harris, 2013), local-field-potential recordings from superficial vs. deep layers (Buffalo et al., 2011; Maier, Adams, Aura, & Leopold, 2010; Maier, Aura, & Leopold, 2011; Spaak, Bonnefond, Maier, Leopold, & Jensen, 2012; Xing, Yeh, Burns, & Shapley, 2012; Bastos, Vezoli, Bosman, Schoffelen, Oostenveld, Dowdall, De Weerd, Kennedy, & Fries, 2015; Michalareas, Vezoli, van Pelt, Schoffelen, Kennedy, & Fries, 2016), and top-down-specific synchronization (von Stein, Chiang, & König, 2000; van Kerkoerle, Self, Dagnino, Gariel-Mathis, Poort, van der Togt, & Roelfsema, 2014). There are a variety of potential mechanisms behind the generation and synchronization of these 5IB bursts (Connors, Gutnick, & Prince, 1982; Lopes da Silva, 1991; Lorincz et al., 2009; Franceschetti et al., 1995; Saalmann et al., 2012). Furthermore, the pulvinar has been shown to drive alpha-frequency synchronization of cortical activity across areas in the alpha band (Saalmann et al., 2012). Behaviorally, there is extensive evidence of alpha-frequency effects on perception consistent with our framework (Nunn & Osselton, 1974; Varela, Toro, John, & Schwartz, 1981; VanRullen & Koch, 2003; Jensen et al., 2012).

The 6CT neurons exhibit regular spiking behavior, in contrast to the 5IB bursting (Thomson, 2010;

Thomson & Lamy, 2007). 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 inputs have metabatropic glutamate receptors (mGluR) that have longer time-scale temporal dynamics consistent with the alpha period (100 msec) and even longer (Sherman, 2014), and significantly more plasticity-inducing NMDA receptors compared to the 5IB projections (Usrey & Sherman, 2018). These properties are both 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 highly localized in time.

The 5IB inputs often have a 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 (Sherman & Guillery, 2006), there is not much evidence of these playing a clear role in the awake, behaving state, and as noted above 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.

Predictions for Predictive Learning

The following are direct, testable predictions from this framework, organized according to a set of different major themes, which also help to clarify the exact nature of the theory in contrast to others.

Learning effects: Measuring effects on learning would be the most direct test of the theory. In the visual domain, this would be most effective by introducing novel sequential "physics" in movies at the alpha time scale, that are inconsistent with standard physics (e.g., reversing the direction of gravity, or having multiple gravitational directions). Some existing research has employed simple stimulus sequences (Gavornik & Bear, 2014; Fiser, Mahringer, Oyibo, Petersen, Leinweber, & Keller, 2016) that are likely learned at a higher, episodic-memory level of encoding, and thus not directly relevant to the lower-level sensory-cortical learning supported by the pulvinar. To distinguish these learning effects from pervasive motor learning, it would be most effective to directly measure activity in the pulvinar and / or associated perceptual neocortical areas, instead of looking at overall behavioral performance.

- Lesioning / inactivating the pulvinar should impair learning. This is perhaps the most obvious prediction, but also challenging to test, for multiple reasons. First, much of the learning in posterior cortex should take place early in development, requiring very early developmental interventions. Indeed, if a primary function of this system is for predictive learning *to train the neocortex*, then once this learning has been achieved, the contributions of this circuit may be much more strongly weighted toward its role in attention, as we discuss below. In any case, it might be difficult to uniquely attribute deficits to learning *per se*, given these additional attentional contributions. Furthermore, existing evidence suggests that inactivation of pulvinar has dramatic effects on cortical activity, raising further interpretational difficulties if learning deficits are seen (Zhou et al., 2016; Purushothaman, Marion, Li, & Casagrande, 2012).

- Blocking synaptic plasticity in pulvinar should have some effect on learning, although most of the

learning occurs in the neocortex as a result of pulvinar broadcasting the temporal difference error signal into neocortex. As noted above, only the 6CT projection should exhibit plasticity effects. And 6CT learning effects are most important very early in the learning process, as the pulvinar learns to map the neocortical representations into the space defined by the 5IB projections. Thus, these effects would require very early interventions.

- Evidence of predictive, anticipatory firing in the pulvinar would constitute an important piece of consitent evidence, but not as causally definitive as direct learning effects. Interestingly, Barczak, O'Connell, McGinnis, Ross, Mowery, Falchier, and Lakatos (2018) recently showed that the auditory pulvinar in monkeys indeed exhibited earlier predictive activity than A1, using a very carefully controlled auditory sequence that had no first-order acoustic differences from a background noise signal. This is suggestive that the pulvinar may be predictively encoding these patterns, but further work would be needed to determine its role in actually learning these sequences in the first place. Note that we would also expect predictive signals in the deep layers of higher auditory areas that contribute to the formation of the pulvinar prediction, but these areas were not recorded in this study. Thus, this paradigm would appear ideal for further tests of this predictive learning framework.

- Another intriguing finding about pulvinar activity demonstrated clear modulation by *confidence* 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 msec of processing, and manifested as a positive correlation with confidence (i.e., more ambiguous stimuli resulted in lower firing rates). We can interpret this as reflecting an ongoing generative prediction of the stimulus signal, with stronger firing associated with more unambiguous predictions based on the current internal representation. Note that this directionality is the opposite of an explicit error-like encoding, 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.

Nature of the Prediction Error Signal: Our thalamocortical framework hypothesizes that prediction error signals are represented as temporal differences between prediction and outcome states, which contrasts directly with the explicit error-coding neurons hypothesized by a range of existing frameworks, most recently that developed by Friston and colleagues (Friston, 2010). As noted above, unambiguous evidence in support of these explicit error-coding neurons is lacking (see Walsh et al., 2020 for a recent comprehensive review). Here, we can articulate the positive predictions of the temporal-difference model, in contrast to the explicit error coding model (EE).

- One of the most consistent positive findings in the literature reviewed by Walsh et al. (2020) is that higher-level activation of predicted stimuli excites associated lower-level representations, in anticipation of stimulus onset. For example, the widely replicated predictive remapping effect is of this nature (Duhamel, Colby, & Goldberg, 1992; Cavanagh, Hunt, Afraz, & Rolfs, 2010). This excitatory nature of prediction / top-down processing is entirely consistent with our framework, in contrast with EE which hypothesizes that it should cancel out (inhibit) consistent bottom-up activation. The key further prediction from our framework is that errors should be encoded as a new state of activity representing the unexpected outcome, immediately following any prior predictions. This prediction has the "unfortunate" status of being overwhelmingly supported by a wide range of existing data, and

also essentially impossible to distinguish from the purely feedforward flow of new sensory inputs. Indeed, it is this "natural" status of the error signal which we find particularly appealing. Nevertheless, in contrast with the EE framework, our model provides a biological mechanism that explains how widely observed predictive learning phenomena can arise without requiring empirically unsupported EE coding neurons.

- Thus, the more distinctive prediction of our framework is that these temporal differences actually drive synaptic plasticity in an error-driven learning manner. 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, and vice-versa if the activity pattern is reversed (LTP for more activity in outcome than prediction). Furthermore, if activity is essentially stable across both phases, then weights should not change. This should be directly testable 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; Whittington & Bogacz, 2019; Lillicrap et al., 2020).

- There is also reasonably strong empirical support for the phenomenon of *expectation suppression*, where expected inputs elicit suppressed neural responses (Summerfield, Trittschuh, Monti, Mesulam, & Egner, 2008; Todorovic, van Ede, Maris, & de Lange, 2011; Meyer & Olson, 2011; Bastos, Usrey, Adams, Mangun, Fries, & Friston, 2012), but multiple comprehensive reviews conclude that it is difficult to distinguish these effects from the neural adaptation effects that underlie the well-documented *repetition suppression* effect (Walsh et al., 2020VinkenVogels17; Kok et al., 2012; Lee & Mumford, 2003). Furthermore, detailed single-neuron level recordings are the least likely to show these effects, which are instead most evident in large aggregate signals such as the BOLD response in fMRI, suggesting that they are likely due to population-level differences in activity, rather than individual explicit error coding neurons. For example, in our framework, unexpected outcomes would be associated with two distinct patterns of activity over a given area (the prediction followed by the outcome), whereas a predicted outcome would only have one, which would be subject to adaptation effects due to the predictive pre-activation. Thus, considerably more detailed and replicable experimental paradigms using single-neuron resolution techniques may be needed to resolve differences at this level.

Alpha Frequency Effects: The 5IB bursting is known to have an intrinsic 100 msec period, which, along with other thalamocortical network effects, defines the predictive learning cycle in our framework. There is a large and growing literature about the behavioral and biological properties of alpha frequency oscillations in thalamocortical networks. Much of this literature is generally consistent with the idea that alpha is important for predictive learning and sensory processing, but also it is clear that there are important attentional modulation effects associated with alpha as well. We will consider these issues later as well, but briefly it is clear that reducing cortical activity and excitability enables intrinsic oscillatory drivers to exert a much stronger synchronizing effect on cortical firing, increasing associated EEG power (Zhou et al., 2016; Klimesch, Sauseng, & Hanslmayr, 2007; Jensen & Mazaheri, 2010). Thus, a reduction of attentional focus brings an associated reduction in activity and excitability, and hence an increase in EEG alpha power. This, however, does not directly disconfirm the idea that predictive learning and attention operate within circuits driven by 5IB alpha-bursting inputs. Instead, it is important to look at other measures such as intertrial phase coherence (ITPC), which more cleanly reflect just the direct effects of these 5IB bursts. Here, the evidence is much more consistent in showing increased ITPC for conditions associated with increased prediction and attention (papers).

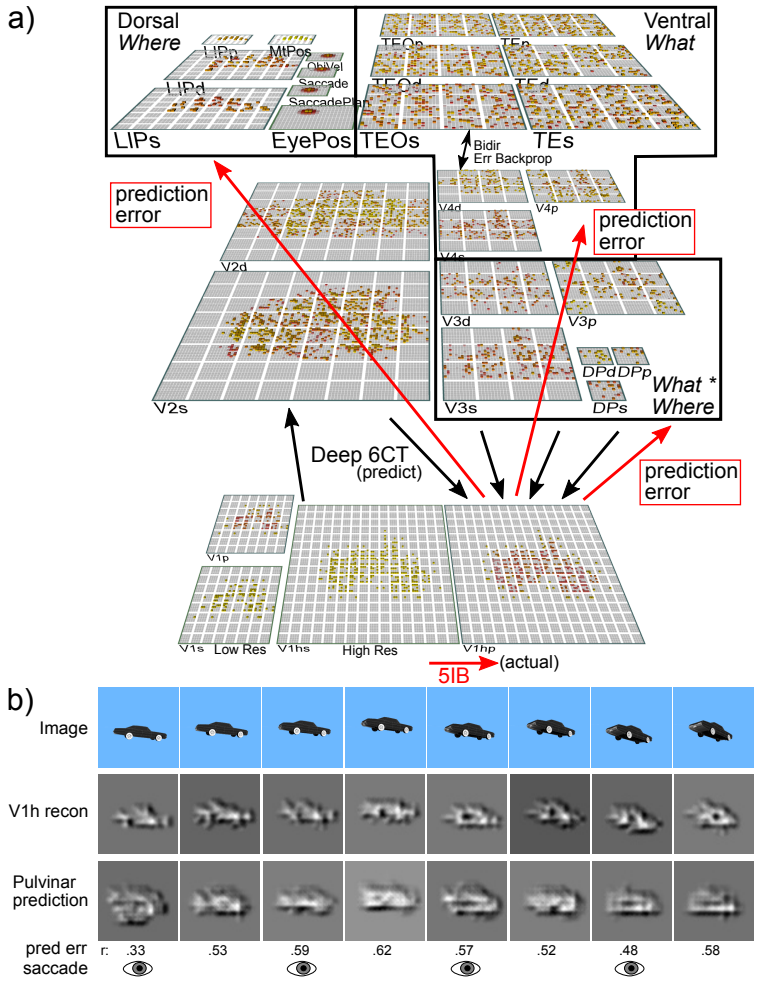

Figure 2: **a)** The *What-Where-Integration, WWI* model. The dorsal *Where* pathway learns first, using easily-abstracted *spatial blobs*, to predict object location based on prior motion, visual motion, and saccade efferent copy signals. This drives strong top-down inputs to lower areas with accurate 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. *s* suffix = 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 (*V1 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 SI for more analysis). Eye icons indicate when a saccade occurred.

Predictive Learning of Object Categories in IT Cortex

A critical question for predictive learning is whether it can develop high-level, abstract ways of representing the raw sensory inputs, while learning from nothing but predicting these low-level visual inputs. For instance, can predictive learning really eliminate the need for human-labeled image datasets where abstract category information is explicitly used to train object recognition models via error-backpropagation? Exist-

Figure 3: **a)** Category similarity structure that developed in the highest layer, TE, of the biologically based predictive learning model, showing *1-correlation* similarity of the TE representation for each 3D object against every other 3D object (156 total objects). Blue cells have high similarity, and model has learned block-diagonal clusters or categories of high-similarity groupings, contrasted against dissimilar off-diagonal other categories. Clustering maximized average *within - between* correlation distance (see SI). 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 1c) to capture coarse perception in model, aggregated by 20 basic-level categories. Each cell is 1 - proportion of time given object pair was rated more similar than another pair (see SI). The human matrix shares the same centroid categorical structure as the model (confirmed by permutation testing and agglomorative cluster analysis, see SI). **c)** Emergence of abstract category structure over the hierarchy of layers. Red line = correlation similarity between the TE similarity matrix (shown in panel a) and all layers; black line shows correlation similarity between V1 against all layers (1 = identical; 0 = orthogonal). Both 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.

ing predictive-learning models based on error backpropagation (Lotter et al., 2016) have not demonstrated the development of abstract, categorical representations. Previous work has shown that predictive learning can be a useful method for pretraining networks that are subsequently trained using human-generated labels, but here we focus on the formation of systematic categories *de-novo*.

To determine if our biologically based predictive learning model (Figure 2b) can naturally form such categorical encodings in the complete absence of external category labels, 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 from the CU3D-100 dataset (O'Reilly, Wyatte, Herd, Mingus, & Jilk, 2013). The objects moved and rotated in 3D space over 8 movie frames, where each frame was sampled at the alpha frequency (Figure 2c). There were also saccadic eye movements every other frame, introducing an additional predictive-learning challenge. An efferent copy signal enabled full prediction of the effects of the eye movement, and allows the model to capture *predictive remapping* (a widely-studied signature of predictive learning in the brain) (Duhamel et al., 1992; Cavanagh et al., 2010), and introduces additional predictive-learning challenge. The only learning signal available to the model was a prediction error generated by the temporal difference between what it predicted to see in the next frame and what was actually seen.

We performed a representational similarity analysis (RSA) on the learned activity patterns at each layer in the model, and found that the highest IT layer (TE) produced a systematic organization of the 156 3D objects into 5 categories (Figure 3a), which visually correspond to the overall shape of the objects (pyramid-shaped, vertically-elongated, round, boxy / square, and horizontally-elongated). This organization of the objects

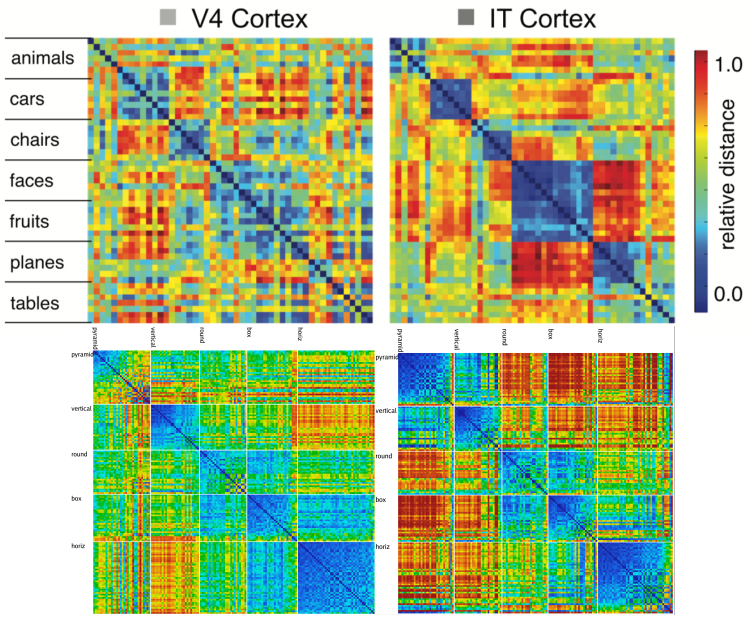

Figure 4: Comparison of progression from V4 to IT in macaque monkey visual cortex (top row, from Cadieu et al., 2014) versus same progression in model (replotted using comparable color scale). Although the underlying categories are different, and the monkeys have a much richer multi-modal experience 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.

matches that produced by humans making shape similarity judgments on the same set of objects, using the V1 reconstruction as shown in Figure 2c to capture the model's coarse-grained perception (Figure 3b; see supporting information for methods and further analysis). Critically, Figure 3c shows that the overall similarity structure present in IT layers (TEO, TE) of the biological model is significantly different from the similarity structure 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. Furthermore, this abstract category organization reflects the overall visual shapes of the objects as judged by human participants, suggesting that the model is extracting geometrical shape information that is invariant to the differences in motion, rotation, and scaling that are present in the V1 visual inputs. 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). This is also evident in Figure 3a by virtue of the close similarity across multiple objects within the same category.

Further evidence for the progressive nature of representation development in our model is shown in Figure 4, which 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. Considerable other work has also compared DCNN representations with these same data from monkeys (Cadieu et al.,

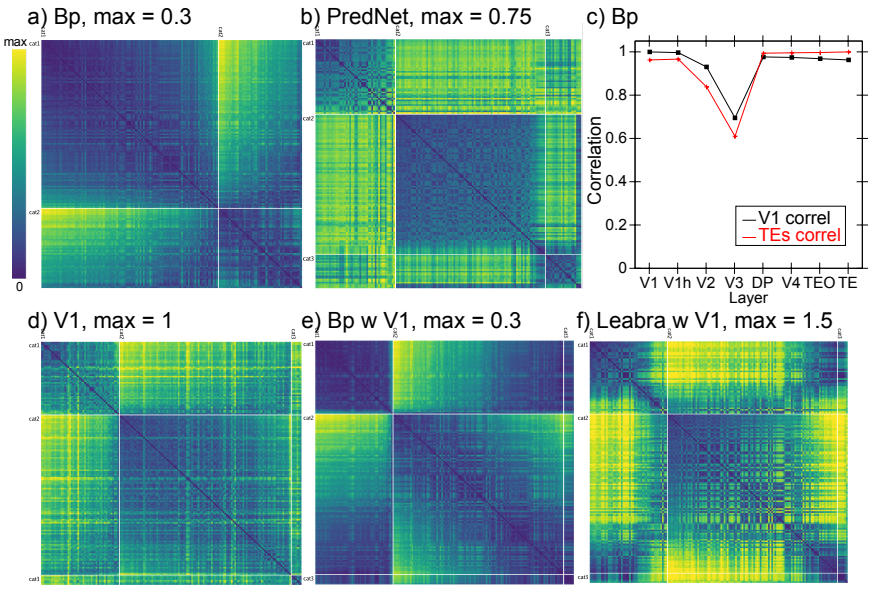

Figure 5: **a)** Best-fitting category similarity for TE layer of the backpropagation (Bp) model with the same What / Where structure as the biological model. Only two broad categories are evident, and the lower *max* distance (0.3 vs. 1.5 in biological model) means that the patterns are highly similar 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)** Comparison of similarity structures across layers in the Bp model (compare to Figure 2c): unlike in the biological model, the V1 structure is largely preserved across layers, and is little different from the structure that best fits the TE layer shown in panel **a**, 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. **d)** 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. **e)** The same similarities shown in panel **a** for Bp TE also fit reasonably well sorted according to the V1 structure (and they have a similar average within - between contrast differences, of 0.0838 and 0.0513 – see SI for details). **f)** The similarity structure from the biological model resorted in the V1 structure does *not* fit well: the blue is not aligned along the block diagonal, and the yellow is not strictly off-diagonal. This is consistent with the large difference in average contrast distance: 0.5071 for the best categories vs. 0.3070 for the V1 categories.

2014), but it is essential to appreciate that those DCNN models were explicitly trained on the category labels, making it somewhat less than surprising that such categorical representations developed. By contrast, we reiterate that our model has discovered its categorical representations entirely on its own, with no explicit categorical inputs or training of any kind.

Figure 5 shows the results from a purely backpropagation-based (Bp) version of the same model architecture, and a standard PredNet model (Lotter et al., 2016) with extensive hyperparameter optimization (see SI). In the Bp model, the highest layers in the network form a simple binary category structure overall, and the detailed item-level similarity structure does not diverge significantly from that present at the lowest V1 inputs, indicating that it has not formed novel systematic structured representations, in contrast to those formed in the biologically based model. Similar results were found in the PredNet model, where the highest layer representations remained very close to the V1 input structure. Thus, it is clear that the additional

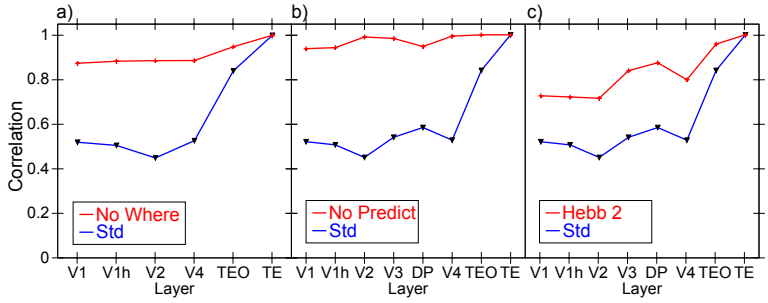

Figure 6: Effects of various manipulations on the extent to which TE representations differentiate from V1. *Std* is the same result shown in Figure 2c from the intact model for ease of comparison. All of the following manipulations significantly impair the development of abstract TE categorical representations (i.e., TE is more similar 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.

biologically derived properties 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. 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 much more computationally intensive and unstable forms of fully recurrent backpropagation (Williams & Zipser, 1992; Pineda, 1987). Furthermore, Hebbian learning requires inhibitory competition which is difficult to incorporate within the backpropagation framework.

Figure 6 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 (Ungerleider & Mishkin, 1982; Goodale & Milner, 1992), 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 6a shows that, indeed, when the *Where* pathway is lesioned, the formation of abstract categorical representations in the intact *What* pathway is significantly impaired. Figure 6b shows that full predictive learning, as compared to just encoding and decoding the current state (which is much easier computationally, and leads to much better overall accuracy), is also critical for the formation of abstract categorical representations — prediction is a "desirable difficulty" (Bjork, 1994). Finally, Figure 6c shows the impact of reducing Hebbian learning, which impairs category learning as expected.

Predictive Saccade-driven Remapping

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 (Duhamel et al., 1992; Colby, Duhamel, & Goldberg, 1997; Gottlieb, Kusunoki, & Goldberg, 1998; Nakamura & Colby, 2002; Marino & Mazer, 2016). 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). 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. Our overall framework 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.

Discussion

In conclusion, we have 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 "deeper" structure of the environment. Relative to existing machine-learning-based approaches in "deep learning", which have generally focused on raw categorization accuracy measures using explicit category labels or other human-labeled inputs, the results here suggest that focusing more on the nature of what is learned in the model might provide a valuable alternative approach. Considerable evidence in cognitive neuroscience suggests that the primary function of the many nested ("deep") layers of neural processing in the neocortex is to *simplify* and aggressively *discard* information (Simons & Rensink, 2005), to produce precisely the kinds of extremely valuable abstractions such as object categories, and, ultimately, symbol-like representations that support high-level cognitive processes such as reasoning and problem-solving (Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005; O'Reilly, Petrov, Cohen, Lebiere, Herd, & Kriete, 2014). Thus, particularly in the domain of predictive or generative learning, the metric of interest should not be the accuracy of prediction itself (which is indeed notably worse in our biologically based model compared to the DCNN-based PredNet and backpropagation models), but rather whether this

learning process results in the formation of simpler, abstract representations of the world that can in turn support higher levels of cognitive function.

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; Olsen, Bortone, Adesnik, & Scanziani, 2012; Bortone et al., 2014; Olsen et al., 2012). The importance of the pulvinar for attentional processing has been widely documented (e.g., LaBerge & Buchsbaum, 1990; Bender & Youakim, 2001; Saalmann et al., 2012), 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; Pinault, 2004; Wimmer, Schmitt, Davidson, Nakajima, Deisseroth, & Halassa, 2015). In our computational framework, these attentional modulation signals cause the iterative 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. A subsequent paper will explore the attentional aspects of the DeepLeabra model and its synergy with the predictive learning aspect.

Synergy between attention and prediction: (Richter & de Lange, 2019) – but in context of expectation suppression as discussed above.

(Kaposvari, Kumar, & Vogels, 2018) (Keller & Mrsic-Flogel, 2018)

TODO: more specific effects.

TODO: Kastner papers. (Halassa & Kastner, 2017) (Fiebelkorn, Pinsk, & Kastner, 2018) (Fiebelkorn & Kastner, 2019)

(Jaramillo, Mejias, & Wang, 2019) – integrative large-scale theory – read fk19a preview first.

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 (Spelke, Breinlinger, Macomber, & Jacobson, 1992; Elman et al., 1996), 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.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*(1), 147–169.

Arcaro, M. J., Pinsk, M. A., & Kastner, S. (2015). The Anatomical and Functional Organization of the Human Visual Pulvinar. *Journal of Neuroscience*, *35*(27), 9848–9871.

Barczak, A., O'Connell, M. N., McGinnis, T., Ross, D., Mowery, T., Falchier, A., & Lakatos, P. (2018). Top-down, contextual entrainment of neuronal oscillations in the auditory thalamocortical circuit. *Proceedings of the National Academy of Sciences*, *115*(32), E7605–E7614.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711.

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J.-M., Oostenveld, R., Dowdall, J. R., De Weerd, P., Kennedy, H., & Fries, P. (2015). Visual Areas Exert Feedforward and Feedback Influences through Distinct Frequency Channels. *Neuron*, *85*(2), 390–401.

Bender, D. B. (1982). Receptive-field properties of neurons in the macaque inferior pulvinar. *Journal of neurophysiology*, *48*.

Bender, D. B., & Youakim, M. (2001). Effect of attentive fixation in macaque thalamus and cortex. *Journal of neurophysiology*, *85*, 219–234.

Bengio, Y., Mesnard, T., Fischer, A., Zhang, S., & Wu, Y. (2017). STDP-compatible approximation of backpropagation in an energy-based model. *Neural Computation*, *29*(3), 555–577.

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *The Journal of Neuroscience*, *2*(2), 32–48.

Bjork, R. A. (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, D. S., Olsen, S. R., & Scanziani, M. (2014). Translaminar inhibitory cells recruited by layer 6 corticothalamic neurons suppress visual cortex. *Neuron*, *82*.

Buffalo, E. A., Fries, P., Landman, R., Buschman, T. J., & Desimone, R. (2011). 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.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology*, *10*(12), e1003963.

Cavanagh, P., Hunt, A. R., Afraz, A., & Rolfs, M. (2010). Visual stability based on remapping of attention pointers. *Trends in Cognitive Sciences*, *14*(4), 147–153.

Colby, C. L., Duhamel, J. R., & Goldberg, M. E. (1997). Visual, presaccadic, and cognitive activation of single neurons in monkey lateral intraparietal area. *Journal of neurophysiology*, *76*, 2841.

Connors, B. W., Gutnick, M. J., & Prince, D. A. (1982). Electrophysiological properties of neocortical neurons in vitro. *Journal of Neurophysiology*, *48*(6), 1302–1320.

Crick, F. (1984). 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.

Dayan, P., Hinton, G. E., Neal, R. N., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, *7*(5), 889–904.

Duhamel, J. R., Colby, C. L., & Goldberg, M. E. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science*, *255*(5040), 90–92.

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, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Fiebelkorn, I. C., & Kastner, S. (2019). A rhythmic theory of attention. *Trends in Cognitive Sciences*, *23*(2), 87–101.

Fiebelkorn, I. C., Pinsk, M. A., & Kastner, S. (2018). A dynamic interplay within the frontoparietal network underlies rhythmic spatial attention. *Neuron*, *99*(4), 842–853.e8.

Fiser, A., Mahringer, D., Oyibo, H. K., Petersen, A. V., Leinweber, M., & Keller, G. B. (2016). Experience-dependent spatial expectations in mouse visual cortex. *Nature Neuroscience*, *19*(12), 1658–1664.

Franceschetti, S., Guatteo, E., Panzica, F., Sancini, G., Wanke, E., & Avanzini, G. (1995). Ionic mechanisms underlying burst firing in pyramidal neurons: Intracellular study in rat sensorimotor cortex. *Brain Research*, *696*(1–2), 127–139.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, *360*(1456), 815–836.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Gavornik, J. P., & Bear, M. F. (2014). Learned spatiotemporal sequence recognition and prediction in primary visual cortex. *Nature Neuroscience*, *17*(5), 732–737.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25.

Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature*, *391*, 481.

Halassa, M. M., & Kastner, S. (2017). Thalamic functions in distributed cognitive control. *Nature Neuroscience*, *20*(12), 1669.

Hinton, G. E., & McClelland, J. L. (1988, January). Learning representations by recirculation. In D. Z. Anderson (Ed.), *Neural Information Processing Systems (NIPS 1987)*, Vol. 0 (pp. 358–366). New York: American Institute of Physics.

Jaramillo, J., Mejias, J. F., & Wang, X.-J. (2019). Engagement of Pulvino-cortical Feedforward and Feedback Pathways in Cognitive Computations. *Neuron*, *101*(2), 321–336.e9.

Jensen, O., Bonnefond, M., & VanRullen, R. (2012). An oscillatory mechanism for prioritizing salient unattended stimuli. *Trends in Cognitive Sciences*, *16*(4), 200–206.

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

Kaposvari, P., Kumar, S., & Vogels, R. (2018). Statistical Learning Signals in Macaque Inferior Temporal Cortex. *Cerebral Cortex*, *28*(1), 250–266.

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

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron*, *100*(2), 424–435.

Klimesch, W., Sauseng, P., & Hanslmayr, S. (2007). EEG alpha oscillations: The inhibition-timing hypothesis. *Brain Research Reviews*, *53*(1), 63–88.

Kok, P., & de Lange, F. P. (2015). Predictive Coding in Sensory Cortex. In *An Introduction to Model-Based Cognitive Neuroscience* (pp. 221–244). Springer, New York, NY.

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron*, *75*(2), 265–270.

Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T., & Miyamoto, A. (2013). Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nature Neuroscience*, *16*(6), 749–755.

LaBerge, D., & Buchsbaum, M. S. (1990). 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.

Larkum, M. E., Zhu, J. J., & Sakmann, B. (1999). A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature*, *398*(6725), 338–341.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America*, *20*(7), 1434–1448.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, 1–12.

Lim, S., McKee, J. L., Woloszyn, L., Amit, Y., Freedman, D. J., Sheinberg, D. L., & Brunel, N. (2015). Inferring learning rules from distributions of firing rates in cortical neurons. *Nature Neuroscience*, *18*(12), 1804–1810.

Lopes da Silva, F. (1991). Neural mechanisms underlying brain waves: From neural membranes to networks. *Electroencephalography and Clinical Neurophysiology*, *79*(2), 81–93.

Lorincz, M. L., Kekesi, K. A., Juhasz, G., Crunelli, V., & Hughes, S. W. (2009). Temporal framing of thalamic relay-mode firing by phasic inhibition during the alpha rhythm. *Neuron*, *63*(5), 683–696.

Lotter, W., Kreiman, G., & Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. *arXiv:1605.08104 [cs, q-bio]*.

Luczak, A., Bartho, P., & Harris, K. D. (2013). Gating of sensory input by spontaneous cortical activity. *The Journal of Neuroscience*, *33*(4), 1684–1695.

Lüscher, C., & Malenka, R. C. (2012). NMDA receptor-dependent long-term potentiation and long-term depression (LTP/LTD). *Cold Spring Harbor Perspectives in Biology*, *4*(6), a005710.

Maier, A., Adams, G. K., Aura, C., & Leopold, D. A. (2010). Distinct Superficial and Deep Laminar Domains of Activity in the Visual Cortex during Rest and Stimulation. *Frontiers in Systems Neuroscience*, *4*(31).

Maier, A., Aura, C. J., & Leopold, D. A. (2011). Infragranular sources of sustained local field potential responses in macaque primary visual cortex. *The Journal of Neuroscience*, *31*(6), 1971–1980.

Marino, A. C., & Mazer, J. A. (2016). Perisaccadic Updating of Visual Representations and Attentional States: Linking Behavior and Neurophysiology. *Frontiers in Systems Neuroscience*, *10*.

Meyer, T., & Olson, C. R. (2011). 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.

Michalareas, G., Vezoli, J., van Pelt, S., Schoffelen, J.-M., Kennedy, H., & Fries, P. (2016). Alpha-Beta and Gamma Rhythms Subserve Feedback and Feedforward Influences among Human Visual Cortical Areas. *Neuron*, *89*(2), 384–397.

Mumford, D. (1991). On the computational architecture of the neocortex. *Biological Cybernetics*, *65*(2), 135–145.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics*, *66*(3), 241–251.

Nakamura, K., & Colby, C. L. (2002). 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.

Neupane, S., Guitton, D., & Pack, C. C. (2016). Two distinct types of remapping in primate cortical area V4. *Nature Communications*, *7*, 10402.

Nunn, C. M. H., & Osselton, J. W. (1974). The Influence of the EEG Alpha Rhythm on the Perception of Visual Stimuli. *Psychophysiology*, *11*(3), 294–303.

O'Herron, P., & von der Heydt, R. (2013). Remapping of border ownership in the visual cortex. *Journal of Neuroscience*, *33*(5), 1964–1974.

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

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, *8*(5), 895–938.

O'Reilly, R. C. (1998). Six Principles for Biologically-Based Computational Models of Cortical Cognition. *Trends in Cognitive Sciences*, *2*(11), 455–462.

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

O'Reilly, R. C., Petrov, A. A., Cohen, J. D., Lebiere, C. J., Herd, S. A., & Kriete, T. (2014). How Limited Systematicity Emerges: A Computational Cognitive Neuroscience Approach. In I. P. Calvo, & J. Symons (Eds.), *The architecture of cognition: Rethinking Fodor and Pylyshyn*[1]*s Systematicity Challenge*. Cambridge, MA: MIT Press.

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. J. (2013). Recurrent Processing during Object Recognition. *Frontiers in Psychology*, *4*(124).

Ouden, H. E. M., Kok, P., & Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology*, *3*(548).

Pennartz, C. M., Dora, S., Muckli, L., & Lorteije, J. A. (2019). Towards a Unified View on Pathways and Functions of Neural Recurrent Processing. *Trends in Neurosciences*.

Petersen, S. E., Robinson, D. L., & Keys, W. (1985). Pulvinar nuclei of the behaving rhesus monkey: Visual responses and their modulation. *Journal of neurophysiology*, *54*.

Petrof, I., Viaene, A. N., & Sherman, S. M. (2012). 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.

Pinault, D. (2004). The thalamic reticular nucleus: Structure, function and concept. *Brain research*, *46*.

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

Purushothaman, G., Marion, R., Li, K., & Casagrande, V. A. (2012). Gating and control of primary visual cortex by pulvinar. *Nature Neuroscience*, *15*(6), 905–912.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). 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.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Richter, D., & de Lange, F. P. (2019). Statistical learning attenuates visual activity only for attended stimuli. *eLife*, *8*, e47869.

Robinson, D. L. (1993). Functional contributions of the primate pulvinar. *Progress in brain research*, *95*.

Rockland, K. S. (1996). Two types of corticopulvinar terminations: Round (type 2) and elongate (type 1). *The Journal of comparative neurology*, *368*, 57–87.

Rockland, K. S. (1998). Convergence and branching patterns of round, type 2 corticopulvinar axons. *The Journal of Comparative Neurology*, *390*(4), 515–536.

Rougier, N. P., Noelle, D., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal Cortex and the Flexibility of Cognitive Control: Rules Without Symbols. *Proceedings of the National Academy of Sciences*, *102*(20), 7338–7343.

Saalmann, Y. B., & Kastner, S. (2011). Cognitive and perceptual functions of the visual thalamus. *Neuron*, *71*(2), 209–223.

Saalmann, Y. B., Pinsk, M. A., Wang, L., Li, X., & Kastner, S. (2012). The pulvinar regulates information transmission between cortical areas based on attention demands. *Science*, *337*(6095), 753–756.

Sherman, S., & Guillery, R. (2006). *Exploring the Thalamus and Its Role in Cortical Function*. Cambridge, MA: MIT Press.

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

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

Sherman, S. M., & Guillery, R. W. (2011). Distinct functions for direct and transthalamic corticocortical connections. *Journal of Neurophysiology*, *106*(3), 1068–1077.

Shipp, S. (2003). The functional logic of cortico-pulvinar connections. *Philosophical Transactions of the Royal Society of London B*, *358*(1438), 1605–1624.

Simons, D. J., & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in cognitive sciences*, *9*(1), 16–20.

Snow, J. C., Allen, H. A., Rafal, R. D., & Humphreys, G. W. (2009). 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.

Spaak, E., Bonnefond, M., Maier, A., Leopold, D. A., & Jensen, O. (2012). Layer-specific entrainment of gamma-band neural activity by the alpha rhythm in monkey visual cortex. *Current Biology*, *22*(24), 2313–2318.

Spelke, E., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of Knowledge. *Psychological Review*, *99*(4), 605–632.

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*(9), 403–409.

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M. M., & Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience*, *11*(9), 1004–1006.

Thomson, A. M. (2010). Neocortical layer 6, a review. *Frontiers in Neuroanatomy*, *4*(13).

Thomson, A. M., & Lamy, C. (2007). Functional maps of neocortical local circuitry. *Frontiers in Neuroscience*, *1*(1), 19–42.

Todorovic, A., van Ede, F., Maris, E., & de Lange, F. P. (2011). Prior Expectation Mediates Neural Adaptation to Repeated Sounds in the Auditory Cortex: An MEG Study. *Journal of Neuroscience*, *31*(25), 9118–9123.

Ungerleider, L. G., & Mishkin, M. (1982). Two Cortical Visual Systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *The Analysis of Visual Behavior* (pp. 549–586). Cambridge, MA: MIT Press.

Urakubo, H., Honda, M., Froemke, R. C., & Kuroda, S. (2008). Requirement of an allosteric kinetics of NMDA receptors for spike timing-dependent plasticity. *The Journal of Neuroscience*, *28*(13), 3310–3323.

Usrey, W. M., & Sherman, S. M. (2018). Corticofugal circuits: Communication lines from the cortex to the rest of the brain. *Journal of Comparative Neurology*, *0*(0).

van Kerkoerle, T., Self, M. W., Dagnino, B., Gariel-Mathis, M.-A., Poort, J., van der Togt, C., & Roelfsema, P. R. (2014). 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.

VanRullen, R., & Koch, C. (2003). Is perception discrete or continuous? *Trends in Cognitive Sciences*, *7*(5), 207–213.

Varela, F. J., Toro, A., John, E. R., & Schwartz, E. L. (1981). Perceptual framing and cortical alpha rhythm. *Neuropsychologia*, *19*(5), 675–686.

von Helmholtz, H. (2013). *Treatise on Physiological Optics, Vol III*. Courier Corporation.

von Stein, A., Chiang, C., & König, P. (2000). Top-down processing mediated by interareal synchronization. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(26), 14748–14753.

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, *1464*(1), 242–268.

Whittington, J. C. R., & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in Cognitive Sciences*, *23*(3), 235–250.

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

Wilson, J. R., Bose, N., Sherman, S. M., & Guillery, R. W. (1984). 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.

Wimmer, R. D., Schmitt, L. I., Davidson, T. J., Nakajima, M., Deisseroth, K., & Halassa, M. M. (2015). Thalamic control of sensory selection in divided attention. *Nature*, *526*(7575), 705–709.

Wurtz, R. H. (2008). Neuronal mechanisms of visual stability. *Vision Research*, *48*(20), 2070–2089.

Xing, D., Yeh, C.-I., Burns, S., & Shapley, R. M. (2012). Laminar analysis of visually evoked activity in the primary visual cortex. *Proceedings of the National Academy of Sciences*, *109*(34), 13871–13876.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–262.

Zhou, H., Schafer, R. J., & Desimone, R. (2016). Pulvinar-cortex interactions in vision and attention. *Neuron*, *89*, 209–220.