Deep Predictive Learning in Neocortex and Pulvinar

Randall C. O'Reilly, Dean R. Wyatte, and John Rohrlich
Department of Psychology and Neuroscience
University of Colorado Boulder
345 UCB
Boulder, CO 80309
randy.oreilly@colorado.edu

October 1, 2018

Supported by: ONR grants N00014-14-1-0670 / N00014-16-1-2128, N00014-15-1-2832, N00014-13-1-0067, D00014-12-C-0638. We thank 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. R. C. O'Reilly is Chief Scientist at eCortex, Inc., which may derive indirect benefit from the work presented here.

Abstract

How does the neocortex learn and develop the foundations of all our high-level cognitive abilities? We present a comprehensive framework spanning biological, computational, and cognitive levels, with a clear theoretical continuity between levels, providing a coherent answer directly supported by extensive data at each level. Learning is based on making predictions about what the senses will report at 100 msec (alpha frequency) intervals, and adapting synaptic weights to improve prediction accuracy. The pulvinar nucleus of the thalamus serves as a projection screen upon which predictions are generated, through deep-layer 6 corticothalamic inputs from multiple brain areas and levels of abstraction. The sparse driving inputs from layer 5 intrinsic bursting neurons provide the target signal, and the temporal difference between it and the prediction reverberates throughout the cortex, driving synaptic changes that approximate error backpropagation, using only local activation signals in equations derived directly from a detailed biophysical model. In vision, predictive learning requires a carefully-organized developmental progression and anatomical organization of three pathways (What, Where, and What * Where), according to two central principles: top-down input from compact, high-level, abstract representations is essential for accurate prediction of low-level sensory inputs; and the collective, low-level prediction error must be progressively and opportunistically partitioned to enable extraction of separable factors that drive the learning of further high-level abstractions. Our model self-organized systematic invariant object representations of 100 different objects from simple movies, accounts for a wide range of data, and makes many testable predictions.

Keywords: Neocortical learning; synaptic plasticity; predictive learning; pulvinar; generative model; vision

Introduction

What is the nature of the remarkable neocortical learning and maturational mechanisms that result in the development of our considerable perceptual and cognitive abilities? In other words, where does our knowledge come from? Phenomenologically, it appears to magically emerge after several months of slobber-filled gaping at the world passing by — what is the magic recipe for extracting high-level knowledge from an ongoing stream of perceptual experience? Answering this central question has been the ultimate goal of many lines of research, at many levels of analysis from synapses up to machine learning algorithms and psychological theories. Despite many advances at each of these levels of analysis, we still lack an overall framework with the key elements of a comprehensive answer to this question: integration across these different levels in a mutually compatible way, with the account at each level having direct empirical support, and directly connecting to the adjacent levels, leaving no obvious theoretical roadblocks. In this paper, such a framework is proposed, providing a broad and deep integration of many different sources of data. This framework is implemented in a computer model that demonstrates both its computational function and its ability to account for a wide range of data.

Our core hypothesis, advanced previously by many different researchers in various forms going back at least to Helmholtz in 1867 (von Helmholtz, 2013), for how substantial amounts of learning can emerge from the largely passive sensory experience of babies, is that each moment can be turned into a *predictive learning* problem. Thus, underlying that passive appearance is an active neural network generating predictions for what will happen next, and learning from the differences between these predictions and what actually does occur. If the brain is actually capable of accurately predicting a reasonable scope of the sensory world, it stands to reason that it must have at least learned a decent internal model of physics, optics, etc, which could thus provide a suitable foundation for higher-level cognitive learning. Within this general framework, several natural questions arise: How frequently are predictions generated and what stimulates their generation? How exactly are the predictions compared with reality, and what form does that critical difference (i.e., the *prediction error*) take, so that it can drive learning? And how can the brain simultaneously represent both a prediction and the sensory ground truth, without getting them mixed up? Finally, what kinds of advanced cognitive functions can be learned through such a predictive learning mechanism, beyond the basic ability to predict subsequent sensory inputs?

Our specific hypotheses for each of these questions are as follows: Predictions (in sensory posterior cortex, at least) are generated every 100 msec (i.e., the alpha rhythm), driven fundamentally by deep layer 5IB intrinsic-bursting neurons which burst at this frequency, entrained via circuits interconnecting the neocorti-

cal deep layers with the higher-order sensory thalamus (the *pulvinar*). Thus, we view this as an automatic, high-frequency, subconscious process specifically for sensory predictive learning — other time scales and forms of predictive learning may occur in other brain areas. In our framework, predictions are generated specifically within the deep neocortical layers (5 and 6), based on time-delayed information from the prior 100 msec, and projected broadly onto the pulvinar thalamic relay cells, via the numerous, weaker "top-down" pathway from neocortical layer 6 (Sherman & Guillery, 2006). After about 75 msec of reflecting these top-down predictions, the sensory bottom-up ground truth drives the pulvinar, via very sparse, strong projections from the 5IB neurons in lower cortical areas (Sherman & Guillery, 2006), and this *temporal difference* reflects the prediction error signal. Thus, unlike most other predictive / generative learning frameworks, we do not propose a population of neurons whose activation explicitly reflects the prediction error — instead the error is implicitly reflected in the temporal dynamics of activation signals emanating from the pulvinar. The pulvinar projects broadly throughout the posterior cortex, and we detail below how this temporal difference at the alpha frequency can drive learning throughout the cortex to improve the accuracy of the predictions generated by the deep layers. Furthermore, while the deep layers are driving their predictions, the superficial neocortical layers are integrating bottom-up and top-down information about the current state of both the environment and the organism, and also learning to improve these representations via the same temporal-difference prediction error signal. Thus, we propose a clear anatomical separation between the predictive (deep layers) and current-time (superficial layers) representations in the cortex — every alpha cycle, the superficial layer state provides the input to the deep layers (again via layer 5IB bursting) that will be used in generating the predictions for the next alpha cycle. Overall, we detail below how this overall framework is consistent with a wide range of established anatomical and physiological data, and furthermore that this predictive learning framework provides a compelling way of unifying all these diverse data under a common function.

As for the question of how far predictive learning can go, we focus on the widely-studied domain of invariant representations of objects that can be used to recognize the specific object being viewed, regardless of where it appears and moves. Such representations are widely recognized as having great adaptive value to an organism, and form the foundation of much of our semantic understanding of the world. However, to develop these representations models typically require training with explicit, invariant category labels due to the strong anti-correlation between the similarity structure at the retinal inputs (where different objects in the same location are more similar than the same object at different locations) and the desired invariant object representations that discriminate between different objects. Thus, if predictive learning can be shown to form such representations in a purely unsupervised manner (i.e., strictly through the process of predicting

subsequent sensory inputs, without any additional high-level category information), then it seems more likely that predictive learning could support a reasonably wide range of higher-level cognitive learning. We explore this question in the context of a simplified, analytic environment where one out of 100 different possible patterns moves in a random direction (or remains still) while the model makes random saccades every 200 msec. This captures the most basic aspects of the visual world: patterns (objects) that are generally stable over time but follow Newton's first law of motion, while also incorporating eye movements, which are the main reliable form of motor control available to a baby. We find that indeed invariant representations do form in the highest layers in the model, corresponding to those in inferotemporal cortex (IT) in the primate brain, and demonstrate that these representations play a critical role in the overall predictive learning process by compactly and stably encoding the visual features present in objects even as the spatial locations where those features appear changes.

Computationally, our framework is a form of a *hierarchical generative model*, which have been widely explored as models of brain / cognitive function (and we restrict our discussion to that subset, broadly defined, as opposed to the broader machine learning field). These models are typically trained progressively from the bottom-up (i.e., layer-by-layer), and according to a relatively strict hierarchy where each layer learns to predict the behavior of the layer below it. We found this approach to have significant limitations, and instead discovered two critical principles that were necessary for the development of systematic, high-level, abstract knowledge representations in our model: 1. Compact, high-level abstract representations are essential for accurate prediction generation at the lowest levels, and thus there must be extensive top-down short-cut projections from the highest levels of the hierarchy down to the lowest levels; and 2. The overall prediction error (broadcast by the pulvinar as a temporal difference) must be progressively and opportunistically partitioned by differentially-specialized such high-level pathways, with simpler factors learned earlier and thus factored out from the overall predictive error signal, thereby concentrating the efficacy of the remaining signal for learning the other aspects of the overall prediction problem. In the case of vision, the spatial (*Where*) aspect of prediction can be learned first, independent of the *What* aspect, facilitated by connectivity that can automatically abstract away feature-level information, leaving only the purely spatial aspect of the signal to be learned. When the accurate spatial predictions from this early-developing *Where* pathway converge on the pulvinar, the residual prediction error signals reflect more of the features present at a given location, rather than where there is visual input overall, and this relative concentration of the error signal then drives the slower-learning *What* pathway to focus its learning on better predicting this featural information, without having to also represent the spatial location of these visual features. In this way, we think this developmental projection of Where-then-What pathway learning is

important for driving the development of invariant object representations, and there is considerable evidence that indeed the dorsal Where pathway does develop first (Bridge, Leopold, & Bourne, 2016).

Although many generative models are discussed in terms of generating predictions, many of them do not actually include an explicit temporal divide, and instead end up learning by reconstructing the *current* sensory input (e.g., an *auto-encoder* in neural network terms). These kinds of auto-encoders require various constraints to avoid degenerate solutions, and it remains unclear whether such models can produce systematic abstract internal representations in a purely self-organizing manner (typically they are subsequently trained with standard explicit object category labels, for example). By contrast, the task of predicting the *future* sensory input avoids many of these problems because, as the saying goes, prediction is difficult, *especially about the future*. We reserve the term *predictive* here exclusively for the *about the future* sense, and discuss the relationship to existing models in detail in the General Discussion section.

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, Colby, & Goldberg, 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, Hunt, Afraz, and Rolfs (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.

This same lesson applies broadly for generating predictions about all aspects of the world, and is why we believe that top-down activation from compact, high-level, abstract representations is essential for the success of predictive learning, and further, that it is consistent with human development of abstract generalization done by refining broad categories rather than building up categories from features. Children, for example, might call all similar four legged animals dog and then refine the concept of dog when the cat's "meow" is not the expected "woof" — i.e., when the actual sound is discrepant with the predicted sound. This discrepancy (prediction error), drives learning at all levels, refining the high level abstract representa-

tion and the more concrete lower level representations as well.

In the following sections, we first provide a more detailed specification of our computational implementation of the predictive learning framework sketched above, along with a discussion of the relevant biological data for each of the major mechanisms in the model, and then test our computational model in the context of visual predictions, to see to what extent it can develop higher-level invariant object representations, in addition to the basic task of predicting what will be seen next.

The DeepLeabra Predictive Learning Framework

In recognition of the critical predictive role of deep neocortical layers, and the ability to train deep hierarchical networks, we refer to our computational model as the *DeepLeabra* learning algorithm, building on our earlier *Leabra* mechanism that performed the same temporal-difference-based error-driven learning in bidirectionally-connected networks, but previously based only on the superficial layers of the neocortex (O'Reilly, Hazy, & Herd, 2016; O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012; O'Reilly & Munakata, 2000; O'Reilly, 1996). A critical feature of Leabra is the ability to effectively and efficiently learn and process information using *bidirectional excitatory connectivity*, which introduces a number of significant computational challenges (but is clearly a major feature of the biology of the neocortex; Rockland & Pandya, 1979; Felleman & Van Essen, 1991; Markov et al., 2014b). In contrast, most existing deep backpropagation models are strictly feedforward, or only do bidirectional processing in a restricted manner. Furthermore, Leabra incorporates both error-driven learning and a robust form of Hebbian learning based on the BCM algorithm (Bienenstock, Cooper, & Munro, 1982; Cooper, Intrator, Blais, & Shouval, 2004; Shouval, Wang, & Wittenberg, 2010), which is essential for successful learning in our model as explored below. Thus, our current model builds directly on this earlier computational infrastructure.

Figure 1 provides an overall schematic for how predictive learning takes place in our framework, showing area V2 predicting the next pattern of activation on V1, over the period of three alpha-cycle "movie frames". The V2 deep-layer neurons drive activation of a minus-phase prediction over the pulvinar, and then in the plus phase the 5IB neurons in area V1 drive the pulvinar with the actual sensory input state, and the temporal difference between the two represents the error signal that trains the superficial and deep layers of V2 to create better representations for making a more accurate prediction next time around. This same cycle of prediction and training occurs for all the layers of the visual system.

The neocortex in our model is composed of two separable but tightly interacting sub-networks, the superficial and the deep / thalamic (pulvinar). The superficial-layer network consists of neocortical layers 4, 2, and

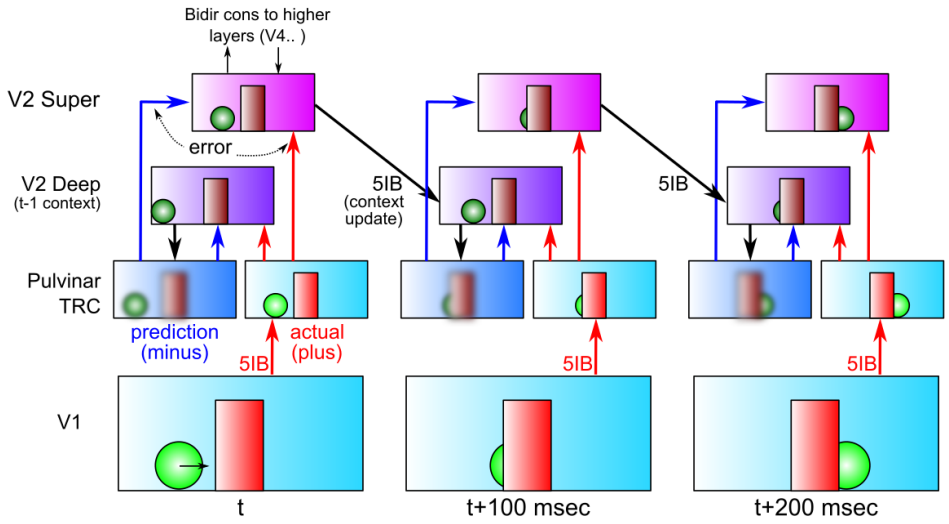

Figure 1: Schematic illustration of the temporal evolution of information flow in a DeepLeabra model predicting visual sequences, over a period of three alpha cycles of 100 msec each. During each alpha cycle, the V2 Deep layer uses the prior 100 msec of context information to generate a prediction or expectation (minus phase) over the pulvinar thalamic relay cell (TRC) units of what will come in next via the 5IB strong driver inputs from V1, which herald the next plus or target phase of learning. Error-driven learning occurs as a function of the temporal difference between the plus and minus activation states, in both superficial and deep networks, via the TRC projections into these networks. The 5IB bursting in V2 drives an update of the local temporal context information in V2, which is used in generating the minus phase in the next alpha cycle, and so on. These same 5IB cells drive a plus phase in higher area TRC's as well, which perform the same kind of *local* predictive auto-encoder learning as shown for V2 here. This system is a predictive auto-encoder (generative model), because it is learning to generate a representation of the V1 inputs (as encoded via the relatively fixed V1 5IB to pulvinar projection).

3, across different brain areas, with extensive bidirectional interconnectivity (feedforward going from 2/3 to layer 4 in the next area, and feedback coming from 2/3 in one area back to 2/3 in an earlier area; Rockland & Pandya, 1979; Felleman & Van Essen, 1991; Markov et al., 2014b). The superficial network represents the current state of the environment and internal state of the organism, at multiple different levels of abstraction, all mutually interacting. It can be described computationally in terms of a classic Hopfield network / Boltzmann machine constraint satisfaction system (Hopfield, 1982, 1984; Ackley, Hinton, & Sejnowski, 1985; Rumelhart & McClelland, 1982), that settles over bidirectional activation propagation updates into a state (representation) that best satisfies the current bottom-up inputs and top-down knowledge / task-driven constraints. The network need not converge fully to a stable settled attractor state — it only moves in that direction within the alpha-cycle time frame, after which changes in the deep / thalamic network (and in the sensory inputs) drive a new settling process under new constraints.

The deep / thalamic network starts in each area with the layer 5b intrinsic bursting (IB) neurons (5IB,Connors, Gutnick, & Prince, 1982; Lopes da Silva, 1991; Sherman & Guillery, 2006; Franceschetti, Guatteo, Panzica, Sancini, Wanke, & Avanzini, 1995; Flint & Connors, 1996; Silva, Amitai, & Connors, 1991), which receive inputs from local superficial neurons and top-down projections from other areas (e.g.,

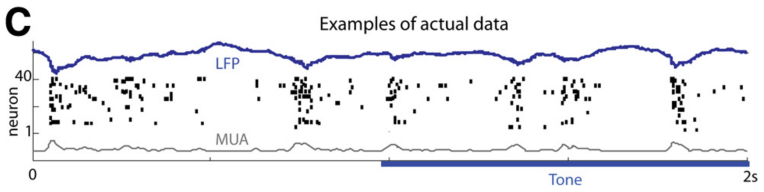

Figure 2: Examples of deep layer 5IB burst firing in primary auditory cortex, in response to a sustained tone stimulus (Figure 1C from Luczak et al, 2013). Dots represent tone-onset aligned spikes, over repeated trials (rows), and top solid blue line represents local field potential (LFP), while bottom grey line represents multi-unit activity (MUA). The phasic bursting of these cells, even in response to a sustained stimulus, is critical for phasically shielding the deep layers from current sensory information (in the long periods after a burst), while also providing a phasic, strong signal to the pulvinar representing the bottom-up ground truth signal, against which the preceding prediction is compared, as shown in the previous figure.

higher-level task control signals). These 5IB neurons then project to deep layer 6, which interconnects with the thalamus (which in turn projects back up to layer 4 of the superficial network and layer 6 in the deep network), and the 5IB neurons also provide a strong driving feedforward input to higher-area thalamic areas. The deep / thalamic network in the posterior cortex is directly responsible for generating predictions over the pulvinar. It must be phasically shielded from the current state information in the superficial layers, to be forced to generate a prediction, as opposed to simply copying the current input state (in which case it would become a simple auto-encoder).

The brief, phasic bursting of the 5IB neurons (with each set of bursts lasting a occurring roughly every 100 msec, see Figure 2 for representative data) is the essential mechanism in our model that ensures that bottom-up, current-state information only penetrates the deep layers phasically, not continuously, thus enabling true predictions to be generated. During the minus phase, when it is generating the next prediction, the deep state reflecting the last 5IB burst of activity is sustained and elaborated through regular spiking layer 6 neurons (i.e., layer 6CT corticothalamic neurons; Thomson, 2010; Thomson & Lamy, 2007) that project to the thalamic relay cells (TRC) of the pulvinar, which then project back to these same 6CT neurons (and up to the layer 4 inputs to the superficial network). Computationally, we divide the 100 msec alpha cycle into 25 msec quarters, with the final quarter corresponding to the time of 5IB bursting and the plus phase (and the prior three quarters constituting the minus phase) — these quarters are thus at the gamma frequency (40 hz), which is typically observed for superficial layer neural firing, and is thought to be modulated by the overall alpha frequency envelope (Dougherty, Cox, Ninomiya, Leopold, & Maier, 2017; van Kerkoerle, Self, Dagnino, Gariel-Mathis, Poort, van der Togt, & Roelfsema, 2014; Haegens, Ncher, Luna, Romo, & Jensen, 2011; Lakatos, Karmos, Mehta, Ulbert, & Schroeder, 2008; Spaak, Bonnefond, Maier, Leopold, & Jensen, 2012; Bollimunta, Mo, Schroeder, & Ding, 2011; Bollimunta, Chen, Schroeder, & Ding, 2008).

Extensive biological evidence supports the alpha-frequency dynamics of the deep layer network (and

gamma for the superficial layers), including direct electrophysiological recording (Luczak, Bartho, & Harris, 2013), local-field-potential recordings from superficial vs. deep layers (Buffalo, Fries, Landman, Buschman, & Desimone, 2011; Maier, Adams, Aura, & Leopold, 2010; Maier, Aura, & Leopold, 2011; Spaak et al., 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 et al., 2014). There are a variety of potential mechanisms behind the generation and synchronization of these 5IB bursts 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, Pinsk, Wang, Li, & Kastner, 2012). Behaviorally, as reviewed below, 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, Bonnefond, & VanRullen, 2012).

Computationally, the deep / thalamic network activations encode temporal context information that reflects activations from the prior 100 msec period, in a manner similar to the simple recurrent network (SRN) model (Elman, 1990, 1991; Jordan, 1989). The SRN is so-named because it employs the *simple* trick of copying the current internal (hidden) layer representation to a context layer that then acts as an additional input to the hidden layer for generating a prediction of what will happen on the next time step. In effect, we hypothesize that the time step for updating an SRN-like context layer is the 100 msec alpha cycle, and during a single alpha cycle, considerable bidirectional constraint satisfaction neural processing is taking place within a DeepLeabra network. This contrasts with the standard SRN, which is typically implemented in a feedforward backpropagation network, where each time step and context update corresponds to a single feedforward activation pass through the network. Our model differs from a standard SRN by pre-computing the context-integrated net input, which deep layer neurons can maintain through bidirectional excitatory loops and longer-lasting channel dynamics, e.g., in NMDA and mGluR receptors. But it fundamentally retains the copy-then-learn dynamic of an SRN, which we argue is essential because subsequent outcomes must be used to determine what is relevant from the past.

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.

Predictive Learning in Visual What and Where Streams

To test the above predictive learning mechanisms, we applied it to a simple visual prediction task with short "movies" of objects undergoing constant self-motion (without acceleration), and randomly directed saccades with an efferent copy of the upcoming saccade motor plan. After the first frame of such a movie, with the benefit of the efferent copy and motion-tuned visual filters, the subsequent frames should be fully predictable, so our first test was whether the model could learn to accurately predict these subsequent frames. However, we were also interested in the extent to which these same predictive learning mechanisms could develop high-level abstract representations of objects that can then provide a more systematic basis for intelligent behavior. For example, by developing invariant object representations, an organism would be able to systematically respond appropriately to the presence of objects regardless of the perceptual details in which that object was viewed.

In developing this model, we explored a wide range of patterns of connectivity, informed by the known biology and also by principles that we developed along the way, that accounted for what worked and what did not (some of which were summarized earlier). Overall, we found a strong correspondence between the successful principles for improving overall network performance, and known features of the biology, reviewed in greater detail below. The extent and depth of this correspondence suggests that structural and developmental properties of the mammalian visual neocortex may have evolved to support the same kinds of computational principles of predictive learning. We first provide an overview of the full model and the simple dynamic visual environment on which it is trained (including saccades), followed by basic computationally-oriented results demonstrating the key principles underlying its learning abilities. Then, we provide detailed accounts of a range of different data of particular relevance to the model, followed by further testable predictions that the model could make.

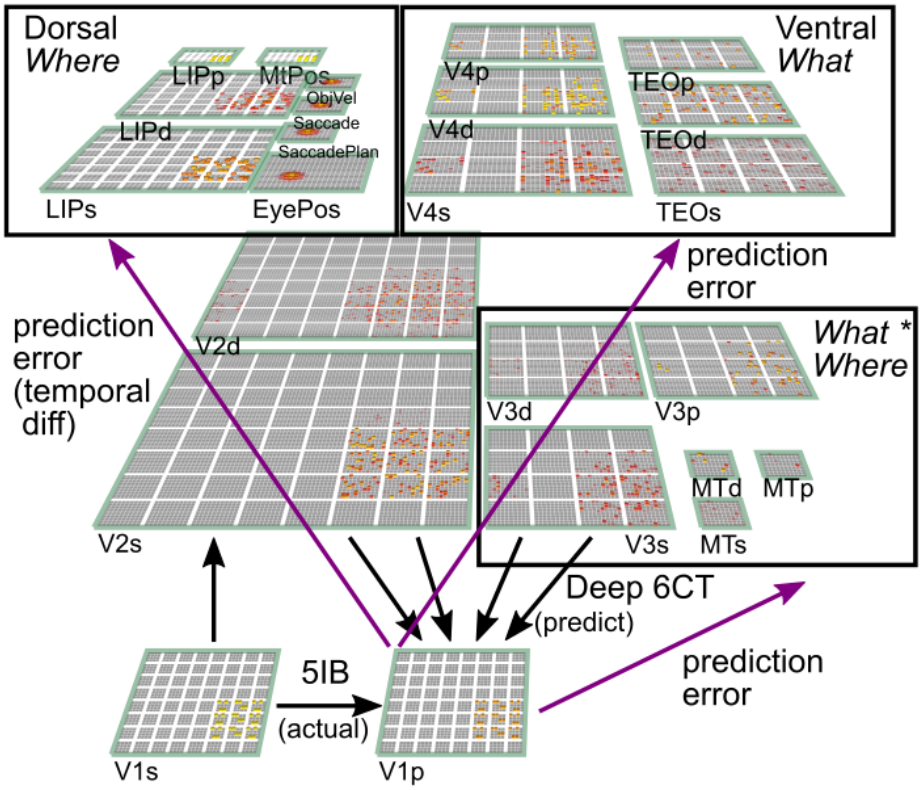

Figure 3: The three-visual-stream deep predictive learning model (What-Where-Integration or WWI model). The dorsal *Where* pathway learns first, using abstracted *spatial blob* representations, to predict where an object will move next, based on prior motion history, visual motion, and saccade efferent copy signals. It then provides strong top-down inputs to lower areas to drive accurate spatial predictions, leaving the residual error to be more about *What* and *What * Where* integration information. The V3 and MT areas constitute the *What * Where* integration pathway, sitting on top of V2 and learning to integrate visual features plus spatial information to accurately drive fully detailed predictions over the V1 pulvinar (V1p) layer (i.e., the cells distributed throughout the pulvinar that receive strong 5IB driver inputs). V4 and TEO are the *What* pathway, and learn abstracted object feature representations, which uniquely generalize to novel objects, and, after some initial learning, drive strong top-down inputs to lower areas. Most of the learning throughout the network is driven by a common predictive error signal encoded via a temporal difference over the pulvinar (V1p and other *p* layers), reflecting the difference between prediction (minus phase) and actual outcome (plus phase). *s* suffix = superficial layer, *d* = deep layer.

The model, which we refer to as the *What-Where-Integration* or *WWI* model, is shown in Figure 3, highlighting the three distinct visual streams (*Where, What,* and *What * Where*) all trained with a strong influence from a common predictive error signal represented as a temporal difference over the pulvinar. The only external inputs to this model are the **V1s** superficial layer activations, reflecting basic feature extraction (e.g., gabor oriented edge filtering) on retinal input signals, the saccade-related signals (anatomically in FEF) of current eye position (**EyePos**), saccade motor plan and efferent copy of last saccade vector (**SaccadePlan, Saccade**), and an object velocity representation reflecting output of known visual motion signals (**ObjVel**) — these last could be directly computed from the V1 inputs but it is simpler to provide as inputs. There is no input of high-level category representations as are typically used in supervised backpropagation networks

— instead this model is entirely self-organizing and forms complex high-level representations without any explicit external shaping forces. We also have a number of *decoders* (not shown in the figure) that receive inputs from various areas in the model, and attempt to decode things like object identity or position — these provide one major means of understanding what these areas are representing (in a manner analogous to typical methods in neuroimaging of the brain). Critically, these decoders do *not* feed back into the network and have absolutely no influence on learning in the model.

According to the known biology of the pulvinar, each of the different areas receives from its own subset of ventral pulvinar TRC neurons, but the wide distribution of V1 5IB driver inputs throughout the ventral pulvinar (Shipp, 2003) suggests that at least a portion of the pulvinar signal shares a common training plus-phase input across all the areas in the model. This 5IB plus-phase input determines the resolution of the prediction that is learned — biologically there may be only a few such 5IB neurons per microcolumn that present a kind of summary output for the entire microcolumn, and we just use a simple one-to-one mapping from our rate-coded microlumn-level superficial layer units. Computationally, it was easier to represent this using a single **V1p** layer that projects to all areas, and also receives deep-layer minus-phase prediction inputs from these same areas, such that predictions reflect the integrated best guesses from different areas and pathways in the model. To measure network learning, we compute the cosine difference between the minus-phase prediction and plus-phase actual input over this V1p layer (cosine is computed as the normalized dot product between the two vectors, separately mean-normalized). The full, trained model produces values around 0.9 or above on this measure, where 1.0 is perfect prediction.

The overall laminar structure and types of connectivity patterns in the model are based on our prior bidirectional object recognition model (O'Reilly, Wyatte, Herd, Mingus, & Jilk, 2013), and follow general biological principles of higher areas being more compact and less retinotopically-distributed than lower layers, using convergent topographic projections to integrate over these lower layers. We did not use any non-biological weight sharing. We extensively explored and optimized layer sizes and connectivity patterns for this model — see Appendix for detailed parameters.

The Dynamic Visual Environment

One critical requirement of a predictive learning model is an environment with sufficiently rich yet predictable dynamics over time to drive interesting learning — one cannot use the kinds of randomly-ordered static images typically used with deep neural networks. The environment model that generates the V1 visual inputs (Figure 4) is designed to capture the most basic and essential features of our physical world: there are spatially contiguous objects with stable visual features over time, that can be moving relative to the observer

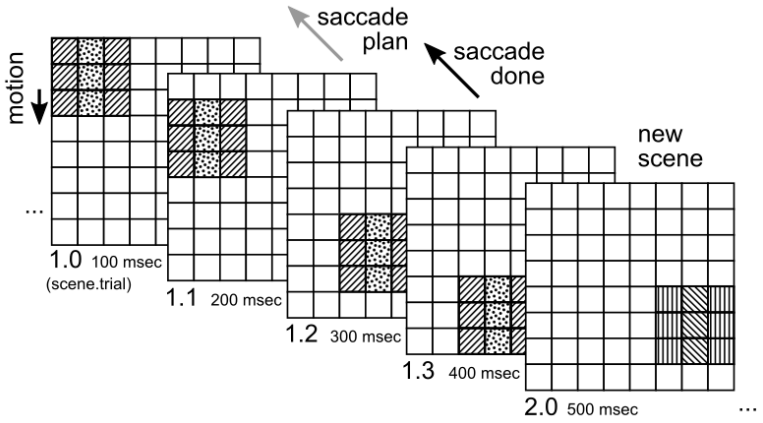

Figure 4: Dynamic visual environment, with 100 different objects composed of two independent sets of features (central column vs lateral flankers, 10 different patterns each), that have a constant motion vector (including the 0,0 no motion case) — a 1 cell per trial downward motion is shown. New scenes are rendered every 4 trials, and each trial represents one alpha cycle (100 msec, 10 Hz). A saccade is planned (i.e., a random vector generated) every 2nd trial, and executed between the 2nd and 3rd trial (note that trial index numbers start at 0). The spatial *Where* pathway can accurately integrate object motion with saccade-generated displacements to predict where the object will appear on the 3rd trial. The *What* pathway can maintain a representation of the object's visual features and apply them consistently across the scene in generating an expectation of what will be seen next. Overall, the model can predict the next trial in this environment with high accuracy (except for the first trial, which is not predictable).

in a stable manner over the period of roughly half a second. Furthermore, the observer can move its eyes in a planned manner (saccades), which results in a discrete displacement of the visual input corresponding to the (opposite) vector of the saccade. Saccades are the main reliable form of motor control that develops first, and including these in the model provides a template for how predictive learning can learn to anticipate the effects of motor actions more generally — it is essential that the visual areas receive information about the *motor plan* (efference copy) in advance of the actual action, to be able to fully anticipate the effects (von Holst, 1954; Wurtz, 2008). This is a form of *forward model* (Kawato, Furukawa, & Suzuki, 1987; Jordan & Rumelhart, 1992; Miall & Wolpert, 1996), as we elaborate in the General Discussion.

To keep things simple and small, we used an 8x8 grid of V1 hypercolumns (each hypercolumn having 4x4=16 feature bits), with an individual object subtending a 3x3 contiguous hypercolumns within that space, without going off the edge. Thus, there are 6x6=36 different locations where the object can appear, and we randomly sampled the motion vector uniformly across the [-2,+2] range of integers (inclusive) separately along the horizontal (x) and vertical (y) dimensions, for a total of 25 different motion vectors. The saccade vectors are drawn from the same distribution. Both such vectors are constrained so as to keep the object fully visible. There is an underlying "world" plane (16x16) where objects are allocentrically located, and eye positions reflect coordinates in this world plane — objects are also constrained to lie entirely within this world plane.

Objects are constructed from two independent sets of features: one for the central vertical column, and

the other for the two flanking columns (Figure 4) . These feature sets comprise 10 random bit patterns with 4 bits active and sharing at most 2 bits with any other such pattern, so there are 10x10=100 total objects under this scheme. We trained the model with 90 of these objects, and reserved 10 for testing. The combinatorial nature of these objects provides a good basis for generalization to the novel testing items. In the real world, the generalization abilities of the human visual system, and large-scale deep neural networks, both support the existence of such a combinatorial (compositional) nature of objects' visual appearance, although the space is certainly much larger and less crisply defined — typical deep neural networks train on 1,000 image categories with roughly 1,000 images per category, and are still likely significantly undersampling the relevant space. Future work will explore scaling up our model to larger, real-object inputs, but the requirement of a dynamic physical simulation for predictive learning makes this much more challenging, as compared to using a large collection of static images. We return to this issue in the discussion.

The temporal structure of the environment is organized into a sequence of *scenes*, with a new scene generated every 4 alpha-cycle *trials*, and a saccade takes place between the 2nd and 3rd trial, as well as between scenes (i.e., after the 4th trial and before the 1st trial of the next scene). The object features remain consistent during a given scene, and change randomly for the next scene. Thus, the first trial is unpredictable, but on the second trial the network has the ability to make an accurate prediction. For this reason, the predictive learning framework requires at least 2 trials (two alpha cycles) for learning to occur. This duration, approximately 200 msec, is consistent with the typical minimum time for fixations (XXX). Recognition can certainly occur in less time but here we are talking about the time needed for the prediction and error signal needed for learning to occur. Having seen the direction of movement on the second trial the network has another opportunity to learn by predicting where to look and what it will see on the 3rd trial and

Another important reason for having 2 such trials is to allow for the planning of a new saccade on the 2nd trial, which is then executed prior to the start of the 3rd trial (i.e., the 3rd trial shows the post-saccade visual inputs). The neural activity representing this planned saccade in the 2nd trial allows the model to accurately predict what the full visual input will be post-saccade. We ignore the actual duration of the saccade, and assume that the system resynchronizes the alpha cycle post-saccade — relevant data are discussed later. There are 2 more trials to process the input post-saccade, and on the 2nd such trial (4th trial of the scene) the model makes a new saccade plan — we assume that even though the object is new, its location is known and so an accurate saccade plan can be generated for the start of the next scene.

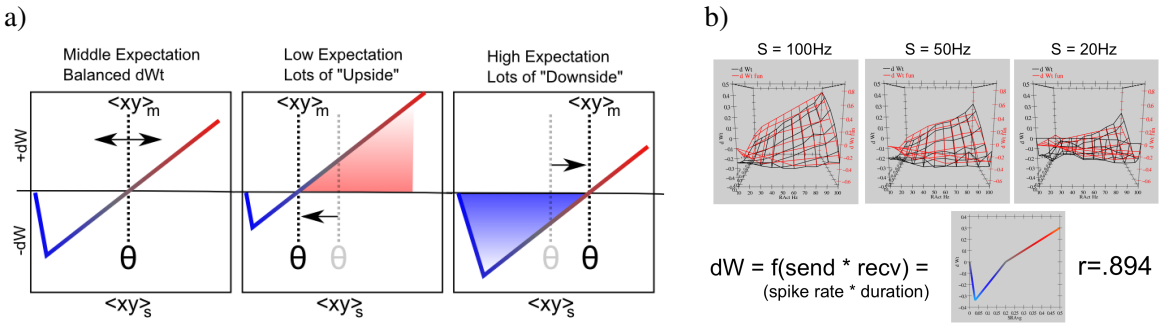

Figure 5: Error-driven synaptic plasticity in Leabra, using the *XCAL* function that is a linearized version of the BCM plasticity function, as derived from the Urakubo et al (2008) STDP model shown in panel (b). a) The threshold θ between weight decrease (-dW, LTD) and weight increase (+dW, LTP) can adapt as a function of recent medium-time-scale average synaptic activity $< xy >_m$, which effectively captures the minus-phase expectation. Learning is driven by the immediate short-term synaptic activity $< xy >_s$, reflecting the plus phase state, and the linear nature of the XCAL function results in an approximation to the CHL equation ($x^+y^+ - x^-y^-$). A more slowly-adapting threshold produces the BCM Hebbian learning dynamics (DROP - featuring a homeostatic negative-feedback mechanism that helps reduce hog units), and a mix of both such learning terms are used. b) The fit to the Urakubo et al (2008) STDP model: a range of sending and receiving spiking frequencies were sampled, and net weight change from the model recorded (black lines). A simple linear equation (the XCAL function) (red lines) fits the overall results well (although the best-fitting function has a small kink around the threshold, a straight line fits nearly as well, and computationally this kink does not affect learning if included).

Model Mechanisms

The model uses standard *Leabra* equations (O'Reilly et al., 2016; O'Reilly et al., 2012; O'Reilly & Munakata, 2000), detailed in the Appendix, for computing rate-coded activation states for each simulated neuron / unit, incorporating both excitatory long-range connections and local inhibitory currents that simulate the effects of inhibitory interneurons. The rate-code activation function closely approximates the well-validated adaptive exponential spiking dynamics of neocortical pyramidal neurons (Brette & Gerstner, 2005), and we assume that an individual simulated neuron in our model corresponds to a population of roughly 100 spiking neurons organized into microcolumns in the neocortex (Buxhoeveden & Casanova, 2002; Mountcastle, 1957, 1997; Rao, Williams, & Goldman-Rakic, 1999). Inhibition is computed as a simple linear proportion of both the *feedforward (FF)* excitatory net inputs to a given area, and the *feedback (FB)* overall activation level within a unit's layer — this *FFFB* inhibition dynamic produces sparse distributed representations within each layer, which have long been shown to be computationally beneficial (Kanerva, 1988; Barlow, 1989; Field, 1994; Olshausen & Field, 1997). Most of the layers have retinotopically-organized hypercolumn-level unit groups within a layer, and the same FFFB inhibitory dynamics operate simultaneously at both the layer and unit group level, with the overall inhibition for a unit being the MAX of each of these computations. This ensures sparse distributed representations both within unit groups and across the entire layer.

Synaptic plasticity in Leabra reflects a synthesis between computational and biological mechanisms. Computationally, it performs both error-driven and Hebbian learning, and we'll see that both of these learning factors are essential for successful learning. The error-driven learning arises from a temporal difference between plus (outcome) and minus (prediction) phases as noted above, approximately of the form of the Contrastive-Hebbian-Learning (CHL; Movellan, 1990) equations:

$$\Delta w \approx \epsilon \left(x^+ y^+ - x^- y^- \right) \tag{1}$$

Where $+$ superscripts indicate plus phase, $-$ minus, and x is the activation of the sending unit, while y is that of the receiving unit. This difference of sender-receiver products computes approximately the same gradient as error backpropagation, subject to symmetry constraints and a few other details (O'Reilly, 1996; Xie & Seung, 2003; Scellier & Bengio, 2017). Critically, each factor in this CHL equation is of a simple xy Hebbian form, making the connection to biological mechanisms more straightforward. We were able to enhance this biological connection significantly by deriving a CHL-like equation directly from a highly detailed biophysical model of spike-timing-dependent-plasticity (STDP; Urakubo, Honda, Froemke, & Kuroda, 2008; Figure 5). Specifically, we found that the rate-code average behavior of this biophysical model, which accounts for a wide range of complex STDP data, can be accurately summarized with a simple linear function that resembles the BCM learning function (Bienenstock et al., 1982; Cooper et al., 2004; Shouval et al., 2010). This function (which we call *XCAL*: temporally eXtended Contrastive Attractor Learning) captures the well-established finding that low (but still elevated) levels of postsynaptic calcium (reflecting the Hebbian xy product) drive a decrease in synaptic weights, while higher levels drive weight increases (Artola, Bröcher, & Singer, 1990; Lisman, 1990, 1995; Bear & Malenka, 1994).

The essential feature of the BCM model is that the threshold crossover point between these two regimes can adapt over time, and by so doing, produce a homeostatic negative feedback mechanism that shifts the balance of weight increases and decreases as a function of how active a unit has been. We realized that if such a threshold were to adapt on a rather more rapid timescale, it could reflect the minus-phase activation state as shown in the CHL equation above, and the linear nature of the learning function then produces the necessary subtraction of this dynamic threshold, with the basic Hebbian-style learning signal reflecting the calcium signal that drives plasticity (Figure 5). Interestingly, some recent data are consistent with more rapidly adapting thresholds (Lim, McKee, Woloszyn, Amit, Freedman, Sheinberg, & Brunel, 2015; Jedlicka, Benuskova, & Abraham, 2015; Zenke, Gerstner, & Ganguli, 2017). Furthermore, our model employs two timescales of threshold adaptation — the shorter one reflecting the minus-phase expectation and a longer

one reflecting overall activation levels over time — thus achieving an elegant synthesis of error-driven and BCM-like Hebbian learning.

In the current model, and most of our other large-scale deep visual models (O'Reilly et al., 2013), the BCM-like Hebbian learning plays a critical role in combating the *hog unit problem*, where a small subset of units takes over much of the representational space and are essentially always active. This problem arises because of the presence of strong positive feedback loops in bidirectionally-connected networks, where units across bidirectionally connected areas can build up mutually reinforcing weights, causing these hogs to form and stabilize themselves. Although error-driven learning should theoretically end up punishing these hog units if they are not contributing to solving the overall problem, it is often the case with challenging problems in deep networks that the error gradients are not very strong or clear at the start of learning, resulting in a kind of "thrashing" dynamic that is ineffective at combating these hog units (and indeed results in a reduction in overall variance in weight values, thereby reducing the random variability that drives exploration of different regions of the solution space). In this context, the BCM Hebbian learning, by raising the learning threshold in proportion to overall unit activation levels, helps to push down the hog units. In addition, we have found that using a normalized momentum learning factor (widely used in backpropagation networks) is helpful for reducing thrashing by driving synaptic weights more quickly along useful gradients, thereby combating hogging as well.

The above mechanisms are used for all neurons in the model, and sufficiently characterize the superficial layers (labeled with an **s** suffix in Figure 3). However, the deep layer and pulvinar neurons have a few special mechanisms to capture their unique functionality. The deep layers in DeepLeabra (with a **d** suffix) capture the firing of the final output stage of the deep neocortical layers, the layer 6CT corticothalamic neurons that project to the pulvinar (and top-down to other neocortical areas) (Thomson, 2010; Thomson & Lamy, 2007). As summarized above, these deep neurons receive a persistent excitatory input representing the SRN-like context information integrated over the superficial layer neurons from the prior alpha trial, and this input is updated as a result of simulated layer 5IB burst firing at the end of every trial. Critically, this prior context state information is the *only* input these deep units receive about the sensory state as represented in the bottom-up feedforward pathways in the network — this restriction is what forces the network to predict, as opposed to simply copy the current sensory input (which is impinging on the superficial layers during the current alpha trial). The V4d and TEOd deep layers also receive a self-context projection, which integrates across the prior deep layer activations in addition to the superficial layers. This supports more enduring activation states over time. We tested this "deeper" context on all layers, but only found benefits for these higher *What* pathway layers, which is consistent with the idea that these areas have more sustained

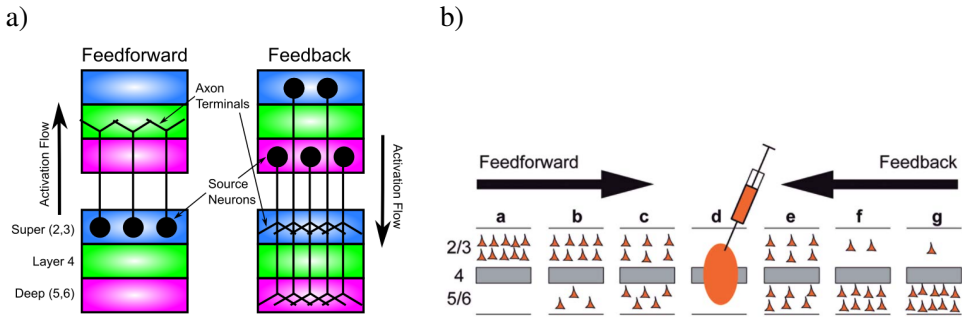

Figure 6: Standard patterns of feedforward and feedback connectivity in neocortex. a) Most feedforward connections originate in superficial layers of lower area, and terminate in layer 4 of higher area. Feedback connections can originate in either superficial or deep layers, and in both cases terminate in both superficial and deep layers of the lower area. (adapted from Fellemen & Van Essen, 1991). b) A more quantitative representation from Markov et al (2014), showing density of *retrograde* labeling from a given injection in a middle-level area (d) — again, 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). However, there appears to be some feedforward contribution from deep-layers, which we did not find to be useful in our model.

representations to support the development of more invariant representations (Foldiak, 1991; O'Reilly & Johnson, 1994; Wiskott & Sejnowski, 2002).

The pulvinar neurons (with a **p** suffix in Figure 3) are specialized to capture the strong driver effects of the 5IB driving inputs — in the plus phase when these neurons fire, their input drowns out the signal from the layer 6CT prediction-generating inputs, and is used as the exclusive source of synaptic input for the pulvinar neurons. Computationally, this is important because simply adding the drivers plus the existing 6CT inputs results in a constantly increasing error signal that drives synaptic weights ever upward (we refer to this as a *main effect* problem). The driving input is computed directly from one-to-one connections from corresponding superficial layer neurons, which are subject to a thresholding process that we assume to be one of the major computational contributions of the 5IB stage.

Connectivity Patterns

Overall, the patterns of interconnectivity among the areas in our model largely follow known biological patterns (Rockland & Pandya, 1979; Felleman & Van Essen, 1991; Markov et al., 2014b; Markov et al., 2014a; Thomson, 2010; Thomson & Lamy, 2007; Schubert et al., 2007; Sherman & Guillery, 2006; Douglas & Martin, 2004), but we also explored many other possibilities, to determine what works best computationally. The resulting model only includes connections with a demonstrated computational value. The computational benefits largely aligned with the known biology. Below, we present results from manipulating a few particularly important connections, which provide key insights into how the model learns.

Starting at the most general level, Figure 6 (adapted from Felleman & Van Essen, 1991) shows that

feedforward connections originate in the superficial layers (2/3) in the lower area, and terminate in layer 4 of the higher area (i.e., the input layer of neocortex, where thalamic inputs from sensory areas terminate in primary sensory areas). From layer 4, connections go straight up to the superficial layers, and in our model we combine the functionality of all of these layers (4,2,3) in the single superficial layer for a given area. Completing the bidirectional loop of excitatory connections within the superficial layers, one type of feedback connectivity originates in the superficial layers of a higher area, and projects back to the superficial layers of a lower area. This pattern of connectivity produces *bidirectional constraint satisfaction* dynamics, iteratively settling into *attractor states* that best represent the constraints present in the external inputs and internal learned synaptic weights (Hopfield, 1982, 1984; Ackley et al., 1985; Rumelhart & McClelland, 1982).

As noted earlier, it is essential in the DeepLeabra model that the feedforward connections *do not* project directly to the deep layers (5,6), because that would give the predictive learning model direct access to the current sensory inputs, which is what it is trying to predict. On the other hand it is very important that the feedback connections from superficial layers *do* drive the deep layers — we found that the deep layers benefit considerably from top-down connections from higher areas, both from other deep layers and from higher-order superficial layers. Computationally, there is the possibility that superficial information from these top-down super-to-deep projections, reflecting current inputs, could short-circuit the predictive learning process. However, because this information is coming only from areas higher in the network, it is already contingent on the quality of the lower-level area in question, and thus is not capable of short-circuiting the learning process. The fact that the deep layers only seem to receive direct feedback is a basic feature of the neocortical connectivity that also makes sense in terms of generative predictive models, where the best source of predictive information comes top-down from, high-level representations.

Figure 7 shows the full pattern of superficial and deep layer connections among all the areas in our model, in comparison to the cortical hierarchy of the macaque from Markov et al. (2014b). For the hierarchically adjacent levels of the *What and What * Where* pathways, the characteristic pattern shown in Figure 6 is present: standard bidirectional excitatory connectivity among superficial neurons, together with top-down projections from both superficial and deep into the deep layers. Note that the top-down and deep-layer connectivity of V1 is omitted from the model, this is only to reduce model running time.

The most interesting connections concern the way that the *What* pathway influences the *What * Where* pathway, which involved the only instances of deep-to-superficial connections (from TEOd to V3s & V2s), in addition to the opposite crossing of superficial-to-deep (from TEOs to V3d & V2d). These connections were essential for allowing more abstract, high-level TEO representations to positively influence the low-

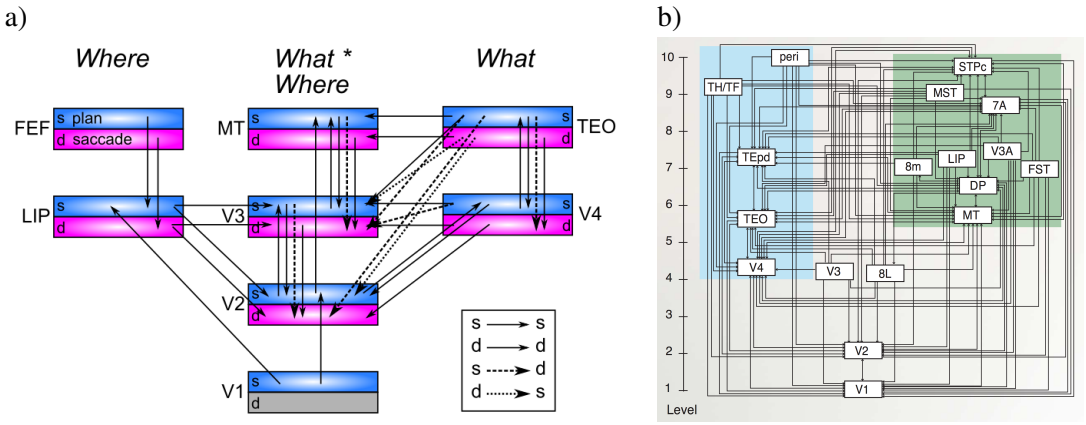

Figure 7: a) Superficial and deep-layer connectivity in the model. Note the repeating motif between hierarchically-adjacent areas, with bidirectional connectivity between superficial layers, and feedback into deep layers from both higher-level superficial and deep layers, according to canonical pattern shown in previous figure. 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 (particularly important for novel test items). b) 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. It is likely that area DP is also part of this integration pathway. 8L = FEF for small-displacement saccades, while 8m = FEF for large-displacement saccades.

level predictions generated over V1p – especially for the novel untrained items.

Next, we consider the interconnectivity with the pulvinar. Biologically, the pulvinar has long remained a bit of a mystery, in part because its obvious anatomical divisions do not appear to coincide with its functional organization — there are coherent retinotopic maps that spread across multiple anatomical divisions, at odd angles, which makes analysis difficult. Shipp (2003) provides an impressive synthesis of the literature, building on the pioneering work of Bender (1981), and clarifies various points of confusion, such that we were able to build our model on the foundation of this synthesis. The major conclusions are that there are four major retinotopically-organized maps in the pulvinar, three corresponding to the ventral cortical pathway, and one for dorsal, and that these maps also have a coarse hierarchical topography, but also considerable levels of intermixing across hierarchical levels.

The first two major ventral pulvinar maps (VP1, VP2) were first characterized by Bender (1981) as being *first-order* and *second-order*, while Shipp (2003) also refers to them as 1° and 2° (confusingly suggesting a difference in visual angle size of receptive field, which is *not* the case). As Bender (1981) emphasizes, these two maps have highly similar properties overall (electrophysiology and patterns of connectivity with cortex), and one primary difference lies in the nature of their topographic organization in the brain, mirroring that of V1 and V2 respectively (where V2/VP2 are wrapped around the central core of V1/VP1). Another major difference is that VP1 (located in inferior pulvinar) receives direct projections from the superior colliculus

which may serve as another source of plus-phase training signal, and could have important implications for spatial attention maps, saccade signals, and also subcortical object / pattern recognition signals (e.g., low-level face detector cells; Morton & Johnson, 1991). For the present model, we use a single common VP substrate. The third ventral pulvinar map, VP3, appears to be dedicated to MT (V5) — we will see below that this may be a separate map because it has a unique developmental trajectory, consistent with the early development of a spatial *Where* system in our model (Bridge et al., 2016). The single dorsal pulvinar map (DP) interconnects with higher-level dorsal pathway areas, including LIP as represented in our model. Shipp (2003) argues that overall the VP3 map can really be considered a part of the DP map — this straddling of ventral and dorsal pathways fits well overall with it playing a key *What* * *Where* integration role in our model.

Our conclusion from this biological data is that the pulvinar serves as a kind of *shared projection screen* (similar to the *blackboard* proposal of Mumford, 1991) where multiple different cortical areas can provide convergent input to shape an overall integrated representation. The projections from pulvinar to cortex then share this converged information broadly back to the same areas that provided input in creating it. As Mumford (1991) emphasized, there is a fundamental puzzle about the pulvinar: it lacks any interconnections among its principal TRC neurons, and therefore does not appear to be capable of doing any processing. This fact is precisely what makes it feasible that it acts as a collection area. Furthermore, the massive projection from pulvinar to cortex, targeting the layer 4 *input* neurons, suggests that the pulvinar is somehow involved in representing the sensory input to the brain. In addition to this projection-screen-like aspect, there is also a rough hierarchical gradient, so the higher-level cortical areas participate more strongly with shaping the more caudal, higher-level representations in the pulvinar, but there is still plenty of mixing here with lower-level cortical areas providing input into these caudal pulvinar areas, and higher-level cortical areas also providing plenty of input into the rostral, lower-level pulvinar areas.

Our model then goes beyond these basic characterizations to further specify that the convergent, integrated representations in the pulvinar are actually *predictions* about what state the strong driving inputs will generate at the next interval of alpha-cycle 5IB burst firing. And the projections from pulvinar back to cortex then carry the critical error signal, in the form of a temporal difference between the prediction and driven states, to train the cortex to produce better such predictions over time. This account helps to make sense of the otherwise somewhat puzzling roles of the two types inputs to the pulvinar (Sherman & Guillery, 2006), and why the strong driver inputs appear to obey the hierarchical topographic organization somewhat more strongly than the other inputs (Rockland, 1998, 1996): this establishes a spectrum of increasingly abstract *ground truth* driver inputs to be predicted. Thus, the "cartoon" of a single projection screen in the pulvinar

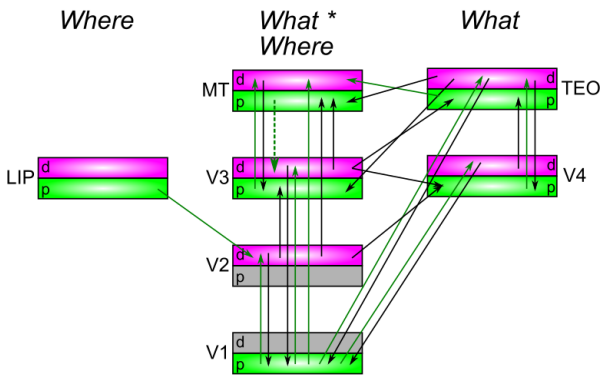

Figure 8: Connectivity for deep layers and pulvinar in the model, which generally mirror the corticocortical pathways (previous figure). Each pulvinar layer (p) receives 5IB driving inputs from the labeled layer (e.g., V1p receives 5IB drivers from V1). In reality these neurons are more distributed throughout 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 MTp $\rightarrow$ V3 and LIPp $\rightarrow$ V2. V2p was omitted because it is largely redundant with V1p in this simple model.

is inaccurate (but a useful first approximation) — it is really a number of different screens at various levels of abstraction.

Figure 8 shows the connectivity of deep layers and pulvinar areas in our model. The overall patterns of connectivity generally mirror those of the corticocortical pathways (Figure 7) — obeying the general *replication principle* of Shipp (2003). Note that the V1d deep layers (6CT) generally project down to the LGN, not the pulvinar, so the next-higher layer, V2d, provides the primary detailed, retinotopically-organized predictive input to the V1p (interestingly, the pulvinar receptive field sizes match those of V2; Bender, 1981). Thus, the extensive top-down corticocortical pathways target V2d, to drive V1p predictions (and we omit V1d from our model). One could label V1p as V2p to align those functions, but there are also distinct pulvinar neurons (anatomically intermixed with V1p neurons) that receive V2 5IB driver inputs, and have similar inputs and outputs as V1p, so we reserve the term V2p for that population of neurons. However, we did not implement V2p in the current model because it was largely redundant with V1p — in the future we plan to add binocular vision and real-world 3D objects, at which point the V2p layer should contain important distinct shape information beyond that in V1p.

The higher-level areas also have their own associated pulvinar layers, which again anatomically are intermixed with V1p, but there is a gradient of the distribution that overall mirrors the caudal-rostral hierarchy of visual areas (Shipp, 2003). These pulvinar layers receive a variety of deep-layer inputs, mostly from neighboring areas, to predict their plus-phase firing patterns. Interestingly, we found a strong constraint on the outputs of these pulvinar areas: they were only beneficial when they projected to higher-level ar-

eas. This makes computational sense in terms of the overall generative, auto-encoder framework, where the higher-level areas are learning to be able to reconstruct lower-level representations. It does not make sense that lower-level areas would have the representational abstractions necessary to accurately drive higher-level representations. Nevertheless, the deep-layer inputs from these lower-level areas can still provide useful information for helping drive the prediction, even though it is not by itself sufficient. This overall constraint is potentially consistent with the patterns of pulvino-cortical connectivity reviewed in Shipp (2003), which appears to be more strongly hierarchically organized compared to the cortico-pulvinar direction. However, more detailed examination of connectivity patterns relative to the strong intermixing of information across the entire pulvinar axis would be necessary to clearly evaluate the validity of this constraint in the biology.

Overall, we argue that the close fit between the characteristic patterns of neocortical / pulvinar connectivity, and the specific, detailed demands of our WWI predictive learning model provides support for the notion that these patterns have evolved to support this functionality.

Early Development of Predictive Spatial Maps in the Where *Pathway*

A central principle of our overall framework is that high-level abstract representations are important for driving lower-level predictions via strong top-down connections. In the case of the dorsal *Where* pathway, it is relatively straightforward to create the relevant spatial abstractions directly from the V1 inputs, and drive predictive learning of object and self-motion (including saccades) on these abstracted spatial *blob* representations at the high levels of the dorsal pathway. The higher levels (e.g., LIP) are compact enough to be capable of remapping saccades over the full span of visual space, whereas in lower levels the degree of interconnectivity across areas would be impossible given the size of the areas. This is consistent with the framework of Cavanagh et al. (2010) (building on Wurtz, 2008), who argue that predictive remapping across saccades is performed at the high levels of the dorsal stream, and it then drives top-down activation in lower areas. Later, we apply our model to account for specific data in the predictive remapping literature.

The two essential features that must be extracted from V1 inputs to make this work are just the retinotopic location irrespective of features (i.e., the spatial blob), and the visual motion vector. Based on a wide range of data discussed next, we hypothesize that area MT (V5) extracts both of these features. The LIP area in our model then integrates these MT inputs together with the saccade plan and actual saccade vector representations (from area FEF and/or superior colliculus) to generate a prediction of where the spatial blob will appear on the next alpha trial, projected onto the LIPp pulvinar. The LIPp is then driven in the plus phase by 5IB bursting output of area MT, providing the ground truth for where the object actually did move.

Due to the relative simplicity of this spatial prediction task, we hypothesized that the brain should learn

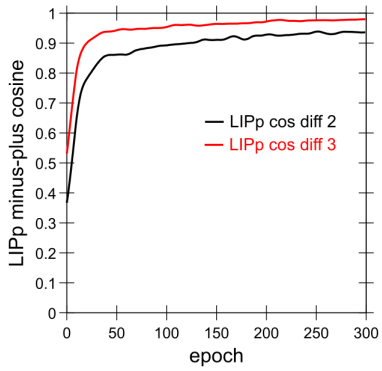

Figure 9: Learning curves for LIP spatial prediction accuracy, measured as cosine between minus and plus phase representations over the LIPp pulvinar layer (perfect accuracy is 1.0). Trial 2 (LIPp cos diff 2, which is the 3rd trial of the sequence) is right after the saccade and thus requires integrating saccade motion plus intrinsic object motion. This curve achieves high levels of predictive accuracy, demonstrating that our model is indeed successfully doing predictive remapping, at least within this *Where* pathway. Trial 3 (LIPp cos diff 3; 4th trial) only requires tracking intrinsic object motion, and is thus easier than the full saccadic remaping task. One epoch = 512 alpha cycles = 51.2 seconds of real time, so this total training period represents approximately 5 hours of real time learning.

it *first*, before anything else of significance is attempted, to absorb as much of the predictive error associated with the spatial aspect, and thereby drive other areas to take on the remaining task of developing representations of *what* is seen. Biologically, this appears to be a well-supported hypothesis. Bridge et al. (2016) review a range of data showing that area MT and its associated VP3 pulvinar area do indeed develop very early, in part through a unique pathway of strong connections from the retina to VP3 (medial inferior pulvinar) that is present early in life, and then is significantly reduced a few months later in development. There is also evidence of direct LGN to MT projections (Sincich, Park, Wohlgemuth, & Horton, 2004). Neurally, area MT matures earlier than other visual areas, at the same time as V1 (Bourne & Rosa, 2006), and behaviorally motion sensitivity develops before form sensitivity in macaques (Kiorpes, Price, Hall-Haro, & Anthony Movshon, 2012). Bridge et al. (2016) also argue that this early development of MT then drives early learning in other dorsal-stream pathways, and that after this early developmental phase, MT shifts over to being driven more strongly by direct V1 inputs and other cortical inputs, as the unique retino-pulvinar pathway retreats.

In our model, we simplify this overall developmental dynamic in several ways. First, we turn off the entire rest of the model for the initial training of the *Where* pathway. Second, we use a separate **MTPos** layer as a proxy for the direct retino-pulvinar pathway, which just collapses all the feature distinctions within a given 8x8 spatial location from the V1 input, producing an entirely spatial input to the LIP. We also use an **ObjVel** input that encodes the visual velocity vector based on object motion, which we assume this early MT layer also provides. Instead of phasing these early drivers out and shifting over to a more

cortically integrated MT later, we just add a new MT layer as shown in the *What * Where* pathway of our model (Figure 3).

We initialized the connectivity of LIP with random weights shaped by topographic sigmoidal and gaussian basis function representations, as has long been recognized as a theoretically-important feature of parietal processing (Zipser & Andersen, 1988; Pouget & Sejnowski, 1997). This improved the learning time compared to purely random weights (see the Appendix for details). The learning curves for this *Where* pathway are shown in Figure 9, for both the post-saccade trial and the trial thereafter. This graph demonstrates that the model is indeed capable of successful predictive remapping using a representation of the saccade plan, integrated with the current object location. Interestingly, as explored later, our model predicts that this predictive remapping happens first in the superficial layers of LIP, and then later and more fully in the deep layers — and these deep layers actually benefit from receiving the actual saccade command, instead of the planning inputs which drive initial updating of the superficial layers. The total training time is approximately 5 hours simulated real-time, with 512 100 msec alpha cycles per epoch, and 300 epochs, which is clearly well within realistic limits. The more complex, higher-resolution learning in the human brain would likely take significantly longer.

Again, we argue that the particular computational demands of our generative predictive learning model align well with the unique developmental trajectory of area MT and associated pulvinar, providing further support for the overall framework.

Later Development of TEO Top-Down Pathway

Another developmental aspect of our model concerns the TEO top-down projections into V3 and V2 — we found small but significant benefits in overall predictive accuracy and ability to decode object information from TEO from delaying the point at which these projections actually influence these lower areas. Computationally, this makes sense because it allows the more fully developed TEO object representations to drive these lower areas, instead of the rapidly changing and initially quite noisy representations. Overall, this reflects an attempt to find a good compromise for the difficult co-dependency problem in the *What* pathway, where high-level abstract representations take a while to develop, and yet are needed for improved prediction performance at the lower levels, which in turn drives better learning of these lower level representations, upon which the TEO representations themselves depend.

Biologically, we were unable to find directly relevant data specifically about the development of top-down projections from TEO, but more general data suggest that IT overall develops relatively slowly compared to other visual areas (Rodman, 1994) and that the visual functions associated with IT emerge rela-

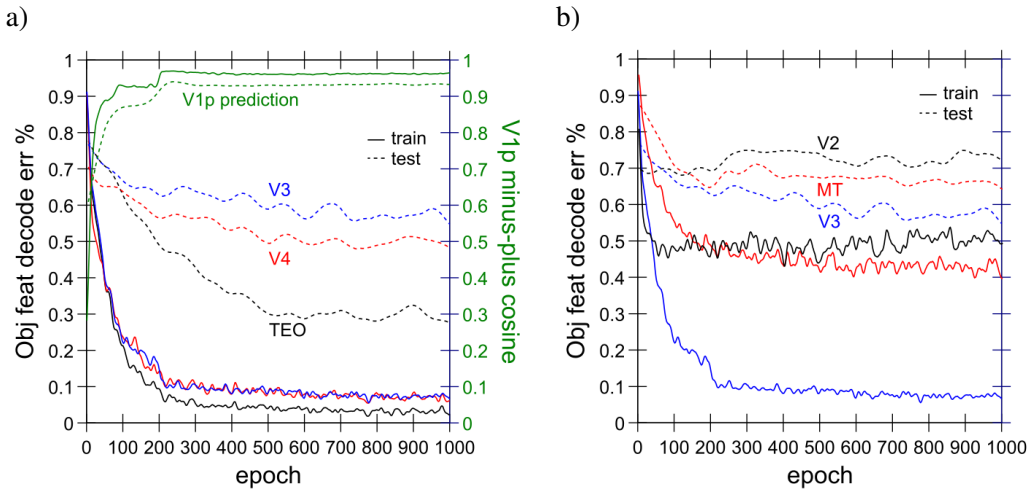

Figure 10: a) Learning curves for full model, showing accuracy (proportion error) in decoding the object features from each of 3 different layers (V3, V4, TEO), and overall prediction accuracy in terms of minus vs. plus phase cosine over the V1p pulvinar layer, at trial 3 (the last trial), which is nearly perfect. Notice that TEO has developed much more systematic object representations than other layers. b) Object feature decoding in layers V2 and MT versus V3, showing that MT specifically seems to learn in the *opposite* direction compared to TEO, producing significantly worse object decoding accuracy compared to V3, which serves as its input. Nevertheless, MT does have slightly better object representations compared to V2. Training curves are bumpier than testing curves because testing occurs only every 5 epochs, and all curves are smoothed with a gaussian filter to remove high-frequency trial-to-trial variance due to differences in environmental inputs. One epoch = 512 alpha cycles = 51.2 seconds of real time, so this total training period represents approximately 16 hours of real time learning. Due to the time required (12 hrs using 64 processors in parallel on our cluster), results are from single runs, but we did run multiple replications of several key conditions and they were very reliable.

tively later in development and continue to develop over a relatively long timecourse (Nishimura, Scherf, & Behrmann, 2009). Thus, this particular feature of our model is overall plausible but not directly supported, and it is quite likely that various other developmental manipulations could have similar benefits, so this remains an area for future exploration.

Results: Understanding how the Model Learns

The first set of results are focused on how the model learns, and how the different pathways and mechanisms interact to produce its overall high levels of predictive learning and development of abstract object representations in the *What* pathway. A second set of results explores how the model accounts for some detailed empirical data of particular relevance.

Decoding Object Features

The learning curves for the full intact model are shown in Figure 10, showing that the model achieves high levels of predictive accuracy in terms of the cosine difference between the minus and plus phase acti-

vation states over the V1p pulvinar layer (green lines, 1.0 is perfect, model achieves roughly .96 on training and .93 on testing). Furthermore, the TEO layer near the top of the *what* stream develops a much more systematic, generalizable representation of objects compared to other layers as it should. This is evident in the ability of the *decoder* (trained using the standard Leabra error-driven learning algorithm, but critically not interacting at all with the model via reciprocal connections) to decode both of the object feature dimensions accurately (each has 10 features, so chance is 1/10 per dimension, or 1/100 for both). The better performance of TEO in decoding is especially evident for the 10 novel testing objects, suggesting that the TEO layer has developed a largely systematic encoding of the object dimensions, supporting roughly 70% accuracy at decoding the object features.

This measure of systematic object feature decoding is not just of computational interest: ecologically, it supports the ability of an organism to accurately and consistently identify objects in the environment, and respond appropriately. Thus, we regard this measure as the most important indicator of overall function in the model: while predictive accuracy is the engine that trains everything, the essential product is the development of a high-level abstract understanding of the environment, that then provides a strong basis for adaptive behavior. Anatomically, TEO provides the input to the higher areas of IT, medial temporal lobe, and ventral and medial prefrontal cortex, all of which build upon these basic invariant object representations to guide goal-driven behavior and high-level memory encoding.

As Figure 10a shows, some of the improved TEO object decoding performance is due to improvements made by V4, indicating the need for multiple processing layers in the *What* pathway, consistent with the biology and recent deep neural network models.

Interestingly, the MT layer shows *worse* object decoding accuracy compared to its input layer, V3 (Figure 10b), indicating that it has learned in the opposite direction from V4 and TEO, in terms of extracting invariant object representations. This oppositional dynamic between MT (i.e., the *What* * *Where* pathway) and IT (the *What* pathway) reflects the critical contributions of the these two pathways in enabling each other to partition distinct parts of the overall prediction problem, and it is evident in many of the other results below.

Decoding Object Position

We examined the ability to decode object position information from various layers, and found that TEO, V4, and MT all had essentially ceiling levels of decoding accuracy. Because we used a gaussian blob spatial representation for spatial location, we measured decoding accuracy in terms of a cosine difference between the target location representation and that produced in the minus phase over the decoding layer (which again

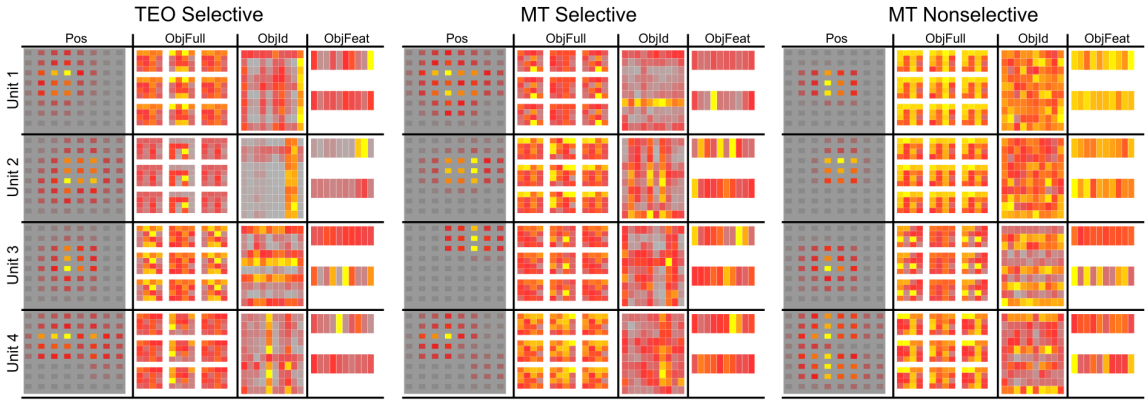

Figure 11: Activation-based receptive fields for TEO vs. MT (superficial layers), selected for relative feature selectivity, and MT non-selective cases. Each cell shows weighted average activation across position and object decoding patterns as a function of unit activity, for 4 target units from each layer (large-scale rows). Pos: position of object, showing large receptive fields in both TEO and MT (the center of the field is sampled more frequently due to nature of sampling constraints, so it is emphasized). ObjFeat: 10 features x 2 dimensions (rows) of the object that was present — e.g., Unit 2 TEO selectively and strongly encodes two of the features from one of the dimensions (top row). ObjId: localist encoding (1 out of 100) of the object identity — due to combinatorial nature of objects, those sharing the same features are aligned vertically or horizontally for the two dimensions, providing a fuller picture of the degree of feature selectivity (i.e., how solid and consistent are the lines). ObjFull: the full rendered object pattern. Overall, TEO has cleaner, more selective ActRF's compared to MT, even in the selected sample (see table 1 for selection details). The non-selective patterns tend to have tighter spatial position coding, and very broad / distributed object coding.

had no interaction with the rest of the network), and these cosines were at 0.995 for these layers for the testing items, and interestingly, somewhat lower for the training items (0.99 for TEO and V4, and 0.98 for MT). Thus, TEO not only encodes abstract object identity, but also spatial location information, consistent with available empirical data (Majaj, Hong, Solomon, & DiCarlo, 2015).

The differences in accuracy between MT and TEO may reflect the comparatively smaller size of MT — when we used a larger MT layer, it started to take on more of the object identity encoding job and this interfered with learning of these object representations in TEO. We hypothesize that the early developmental engagement of the MT more strongly biases it toward spatial representations, which could have the same overall effect as constraining its size as we do here.

Nature of TEO vs. MT Representations

To better understand the nature of the representations that developed in the high layers of the model, we used a form of the *spike triggered averaging* technique that computes a weighted average of the activation state across the network, weighted by the activation level of a given *target* unit (we refer to this as an activation-based receptive field, or ActRF). When the target unit is off, then those network states are effectively ignored (they are multiplied by 0 in the weighted average). And to the extent it is on, the result is an average, weighted by strength of activation, of the activation states correlated with the activity of the

		Spatial RF Size		Cos Trial 2-3 Consistency	
Area	**% Selective**	**All**	**Selective**	**All**	**Selective**
TEO	60%	64%	71%	0.73	0.80
MT	30%	57%	67%	0.60	0.71

Table 1: Quantitative analysis of selectivity, stability, and receptive field size for ActRF representations in TEO vs. MT. Selectivity was cheaply determined by thresholding average activation in the ObjId ActRF — by experimentation, a threshold of 0.4 (on max-normalized 0-1 data) did a good job of separating the feature-selective (having clear lines in the Id ActRF) vs. more complex non-selective units. There were twice as many such selective units (% Selective) in TEO compared to MT, and the majority of TEO units were selective. The next two columns show the average percent of object position cells that units responded to, for All units and for the selectively responding ones, showing that the feature-selective units had larger receptive fields, and that these fields on average covered a large portion of the spatial locations. MT receptive fields were smaller overall. The final two columns measure the consistency (cosine similarity) of the ActRF's computed on trial 2 (immediately post-saccade) vs. trial 3 — the selective ones are more consistent across time, and TEO is more consistent than MT over time.

target unit. In other words, it gives you a pretty clear picture of what the activation patterns in the rest of the network are like when this unit is responding. Furthermore, it can be used with any kind of pattern, even ones not directly connected to the target unit — including the decoder patterns which provide a very clear analysis of the unit's response profiles.

Figure 11 shows the ActRF patterns for a sample of more feature-selective TEO and MT units, and non-selective MT units (which were a majority in MT, while the feature-selective ones were a majority in TEO; Table 1). As explained in the figure, the object ID and feature decoder layers allow us to see how consistently the TEO units respond to a subset of feature values, across a range of different spatial locations. This clearly shows that TEO units have developed the characteristic invariant object recognition property of actual TEO neurons, responding systematically to subsets of object features across a range of locations. Table 1 shows that 60% of the TEO units had this object-feature selectivity, while only 30% of MT neurons did (and even with those, the tuning was less clear and consistent than in TEO). This table also shows the percent of all 64 spatial locations where units responded, showing that TEO had larger receptive fields than MT, and that the feature-selective receptive fields are larger on average than the non-feature-selective ones.

The non-feature-selective receptive fields in MT and TEO (Figure 11) tended to have more focal spatial coding, and broader distributed object feature tuning (including cases with essentially no feature selectivity at all). These are clearly going to be more useful for the *What * Where* integration process, and their prevalence in MT supports this functional role for this area. Nevertheless, these unit types also developed in TEO — as is typical in neural network models, and in the brain, a full distributed spectrum of neural coding types tend to emerge over learning across all areas — there are no truly representationally *pure* areas (Behrmann & Plaut, 2013). This is overall consistent with available data on TEO neurons, which also encode spatial location along with many other properties, and have a broad range of selectivities (e.g., Hong, Yamins, Majaj, & DiCarlo, 2016; Majaj et al., 2015; Zoccolan, Kouh, Poggio, & DiCarlo, 2007; Tanaka, 1996; Logothetis &

Sheinberg, 1996). More generally, these results are consistent with coarse-coded distributed representations of high-dimensional data (also known as *mixed selectivity* Fusi, Miller, & Rigotti, 2016), which are useful for efficiently binding multiple features into a coherent object representation (Hinton, McClelland, & Rumelhart, 1986; O'Reilly & Busby, 2002; O'Reilly, Busby, & Soto, 2003; Cer & O'Reilly, 2006). The greater complexity and higher-dimensionality of the *What * Where* pathway reflects their particular specialization for this kind of binding, but the differences are clearly quantitative, not qualitative.

One further analysis we performed was to compare the consistency (cosine similarity) of ActRF patterns based on activity on trial 2 (immediately post-saccade) to those from trial 3. This provides an indication of how temporally stable these representations are over the 4 trial scene where a single object is present. Table 1 shows that again TEO had overall more such consistency compared to MT, and that the feature-selective units were more consistent than the non-selective ones.

Taken together these analyses strongly show that, consistent with the decoding results, the model's TEO has developed systematic invariant object representations, without any external pressure to do so. This purely self-organized learning, in an environment with a relatively large number (100) of highly overlapping and confusable objects, goes beyond existing auto-encoder neural network models, that tend to extract broad central tendencies across the inputs (e.g., the famous Google auto-encoder network that extracted a blurry cat face from millions of images from the internet; Le, Monga, Devin, Chen, Corrado, Dean, & Ng, 2012). Success in these auto-encoder models is instead typically measured in terms of reductions in number of supervised training trials required on top of the auto-encoding pre-training (Valpola, 2014; Rasmus, Berglund, Honkala, Valpola, & Raiko, 2015).

Importance of a Deep Hierarchy: Testing Flatter Models

Figure 12a shows the effects of removing the higher levels of the network, demonstrating that a deep hierarchy of layers is important for achieving high levels of predictive accuracy in this task, particularly with respect to the novel test items. Performance on these test items indicates to what extent the model is shaping predictive mappings specifically around the trained objects (resulting in poor testing performance), versus having a more generalized, abstract capability of mixing independent *What* and *Where* pathway information (resulting in good testing performance). With only V2, prediction accuracy on V1p is dramatically worse, with cosine levels between .3 and .4 and not much sign of learning progress overall. Adding V3 improves training performance dramatically — the more compact representations and integrative connectivity of V3 adds considerably more systematicity and power. Nevertheless, the performance on the testing items remains differentially lower compared to the training performance, suggesting that the V3-only network is missing

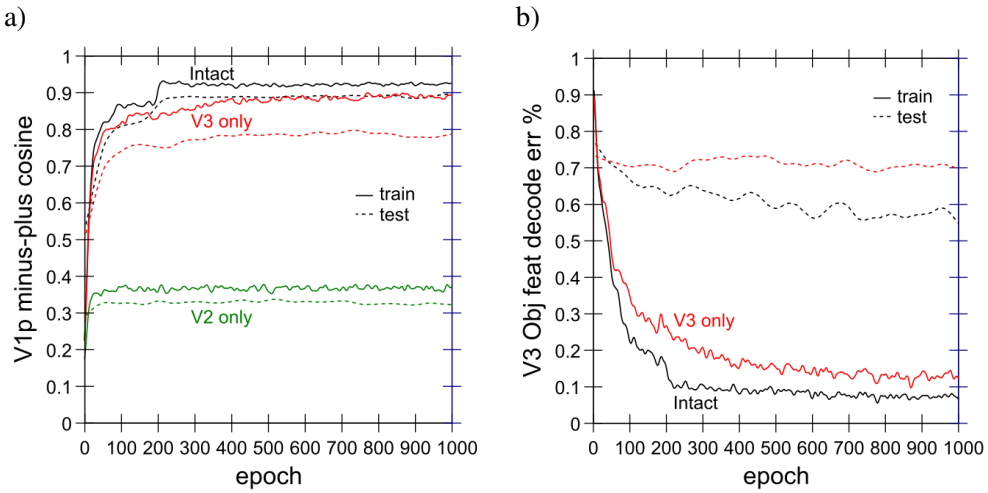

Figure 12: a) Prediction accuracy (minus vs. plus phase cosine over the V1p pulvinar layer), at trial 2 (the post-saccade trial) for model with only V2 (no V3, MT, V4, TEO) or only V3, compared to the full Intact model. A single layer alone (V2 only) cannot do very well, despite getting nearly-perfect spatial inputs from the pre-trained LIP *Where* network. Adding V3 on top of V2 produces a dramatic improvement, but the novel testing patterns are notably worse than the trained ones. b) Object feature decoding accuracy from layer V3 in V3 only vs. Intact model, showing that the top-down projections from higher layers play a significant role in shaping the object encoding in V3 in the Intact model.

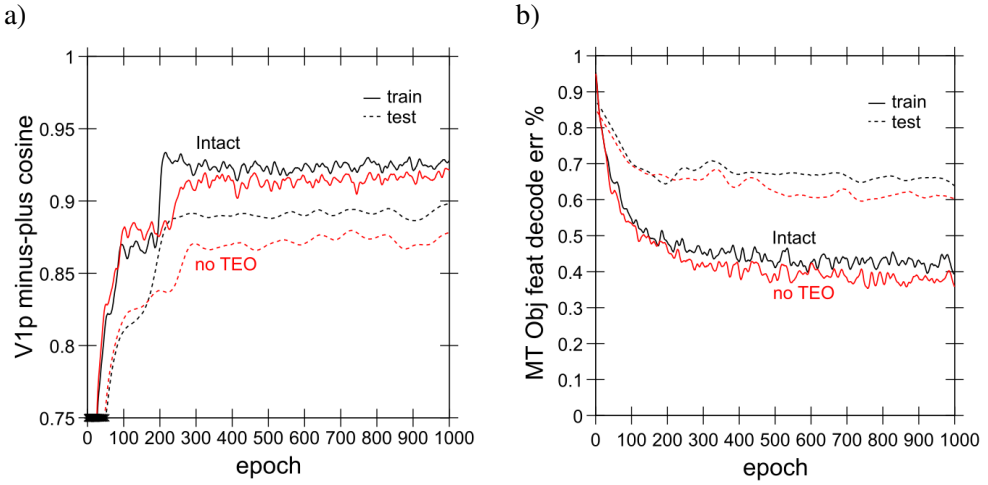

Figure 13: a) Prediction accuracy (as in prior figure) for Intact versus model with no TEO area, showing small but reliable impairment, more for test than trained objects. b) Object feature decoding accuracy from layer MT for Intact vs. no TEO model, showing *improvement* in object detection in MT when TEO is lesioned, consistent with opponent relationship between these pathways.

the ability to more systematically represent objects. Figure 12b reinforces the importance of yet higher layers above V3: these higher layers (MT, V4, TEO) provide a top-down shaping influence on the V3 representations that makes it easier to decode the object features from V3.

Figure 13 shows effects of only removing the TEO area, with everything else as in the full Intact model. This results in a small but reliable impairment in prediction accuracy, more for the novel testing objects than

the trained objects, consistent with the importance of the abstract high-level TEO representations providing top-down drive into the lower-layer predictions. The bump up at 200 epochs, seen in Figure 13a with the constrained vertical scale is when the top-down TEO connections are turned on (see Figure 14b for a direct comparison with top-down TEO connections on starting at epoch 0) reflecting the hypothesized delay in maturation of these projections. The no-TEO model also shows a bump, but at epoch 250, which is when we drop the learning rate on our standard learning rate schedule, which overall produces better learning results and reflects a likely developmental slowing of effective learning rate. Overall, we anticipate that with more complex, high-dimensional real-world objects, this high-level TEO contribution to overall prediction accuracy will be significantly more important, compared to the relatively simple objects used here. Nevertheless, even in this simple case, and especially in the novel testing objects, we obtain an indication of these top-down effects.

Another manifestation of the opponent-dynamics between MT and TEO is evident in Figure 13b, showing the object decoding accuracy in area MT for both the Intact and no-TEO models. Interestingly, the ability to decode objects actually *improves* in MT with the TEO removal, suggesting that it is partially taking on some of the *What* pathway function that TEO otherwise dominates in the intact model. We also tested the removal of MT — in earlier versions of the model this consistently produced major reciprocal impairments on object encoding in TEO, as TEO took on more of the *What* * *Where* integration task from the missing MT. However, due to various improvements in the V4/TEO pathway parameters, it became more robust and the removal of MT only had relatively small (but reliable) effects on TEO object decoding (not graphed).

Developmental Timing: Early Where *and Late* What *Pathways*

The importance of the early development of the LIP spatial prediction pathway on subsequent learning in the full network is shown in Figure 14a. The main effects from not using the pretrained LIP pathway weights are on the development of systematic object feature representations in TEO, reflected in significant reduction in object decoding accuracy on test items, and a corresponding impact on V1p prediction accuracy specifically for these test items. The relatively large impact on testing object decoding is interesting given that the LIP trains quite quickly (a majority of the learning takes place within the first 10 epochs; Figure 9). This again suggests that the partitioning of the spatial component of prediction error is important for allowing the TEO to develop more systematic object encodings, and that doing so before the TEO has any significant learning pressure is critical.

Figure 14b shows the advantages of a developmental delay in the strengthening of the top-down projections from TEO to lower areas (V2, V3). By waiting until the TEO area has had a chance to develop more

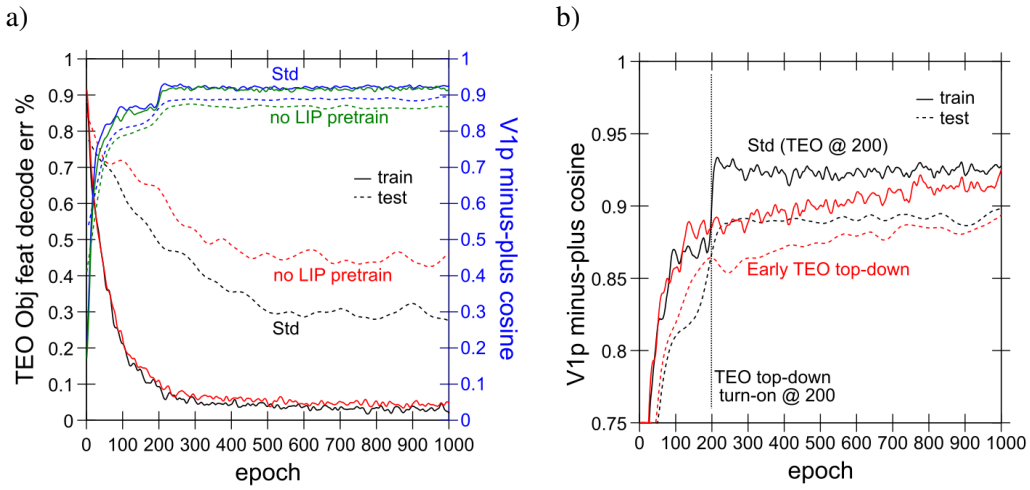

Figure 14: a) Learning without first pretraining the *Where* LIP pathway compared to the standard (Std) training, this has a significant impact on the development of systematic TEO object representations, particularly for the testing items. This has corresponding effects on V1p prediction accuracy (top lines), again particularly on the testing items (the size of these effects is roughly proportional to the relatively small overall impact of TEO on prediction error as shown in earlier figures). Overall, this again supports the importance of partitioning the prediction error so that the TEO can focus on learning more directly about object features. b) Prediction accuracy effects of having top-down TEO to V2,V3 projections effective right from the start of learning, as opposed to coming on after 200 epochs as in the standard model. The delayed engagement of TEO allows overall predictive performance to improve significantly earlier.

abstract object representations, the impact of these more systematic representations produces an immediate bump in predictive accuracy, whereas when these lower layers have first learned to incorporate the less systematic initial TEO representations, it takes much longer to overcome that initial learning and begin to incorporate the more systematic top-down inputs.

Limitations of Outside-In Progressive Learning

Next we tested the standard approach of training deep hierarchical auto-encoders and related models, where progressively higher layers are added after earlier layers have had a chance to develop their initial representations. We did this by using the weights from the V2 only and V3 only cases described above as initial starting weights for training the full standard model. Figure 15 shows that this significantly impaired the ability to decode object features from the TEO area of the model. We argue that this resulted from these models developing representations that tried to solve all aspects of the prediction problem without the benefits of more abstract higher-level representations driving top-down input into these lower layers. Interestingly, the V3-only case was significantly worse here compared to the V2-only, even though V3-only did a better job overall of prediction (Figure 12). This suggests that the representations developed during this initial pretraining fused the *What* and *Where* aspects of the prediction problem in a way that made it difficult

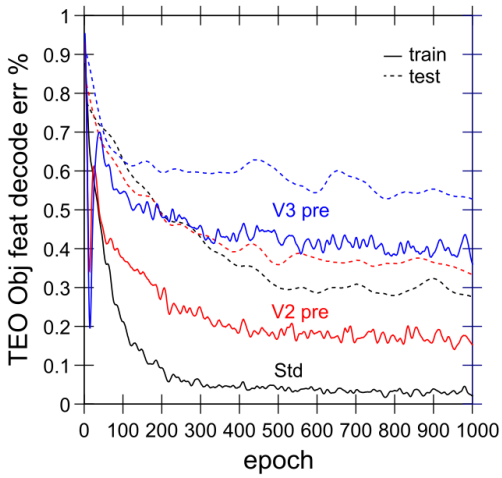

Figure 15: Effects of pretraining using weights from V2 only or V3 only model on object decoding accuracy from the TEO area, as a test of the standard outside-in developmental training approach. This significantly impairs the development of systematic invariant object representations in TEO, presumably by interfering with the prediction error partitioning process, and the top-down influence of more abstract object representations during learning.

to then extract a more pure object-invariant representation. Instead, we argue that our standard version of the model depends critically on the interactions between MT and TEO pathways *from the very start of the learning process* for partitioning the prediction problem, allowing TEO to more fully develop its more pure *What* representations.

Also, these pretrained models actually did relatively well at the V1p prediction learning task, with the V2-pre case even doing slightly better than the default model, suggesting that prediction error in this simple model may not fully reflect the beneficial contributions from high-level abstract representations. We anticipate that with more complex, high-dimensional real-world objects, these high-level representations will be essential for accurate prediction.

Importance of V1p for Higher Areas

One of the potentially puzzling aspects of the pulvinar connectivity is that it appears to route information from low levels of the visual hierarchy (V1, V2) into the higher-level areas such as V4 and TEO. How could such a low-level signal, reflecting detailed prediction errors in our model, be beneficial for shaping higher-level representations? As we have argued above, we think this signal is useful in the context of interactions with other areas, to help partition the overall prediction error signal, such that the *What* pathway ends up being able to focus on improving the prediction accuracy specifically for the object features component. In other words, this shared projection-screen-like representation enables the different areas to effectively coordinate and specialize on specific aspects of the overall prediction task. Throughout the development of

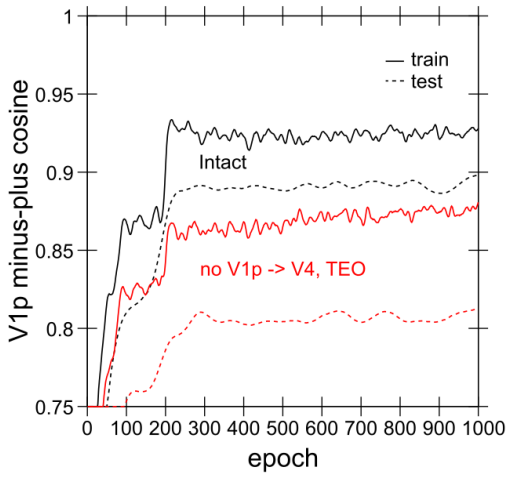

Figure 16: Effects of removing the V1p to V4,TEO projection on overall V1p prediction accuracy, showing similar effects to a TEO lesion, indicating that the *What* pathway is essentially non-functional without this V1p pulvinar projection. Consistent with this, object decoding accuracy in TEO was also completely abolished (not shown).

our model, we consistently found that removing the V1p projections to TEO or V4 impaired performance (object decoding and prediction error) significantly. And in the final model, removing this projection from *both* V4 and TEO results in a *complete failure* to be able to decode object features from TEO or V4. These layers instead develop some entirely different form of representations, and prediction accuracy also suffers significantly (Figure 16). However, there are only relatively minimal effects in the final model of only removing V1p projections to TEO, increasing the object decoding error for trained objects from around .05 to .1, and, surprisingly, having no effect on test objects. Thus, we think that TEO can largely receive the relevant V1p error signals indirectly through its interconnections with V4, but removing this signal from both V4 and TEO is catastrophic.

Also, it is worth noting that throughout most of our model development, we had a small bug in the environment program, which resulted in occasionally unpredictable input sequences being presented. It is possible that the magnified effects of the V1p to TEO projection in these earlier models may reflect its importance for more robust, fault-tolerant learning. We plan to explore this idea in future research.

Importance of Temporal Context, Hebbian Learning, Momentum

Finally, we report the effects of various important elements of the DeepLeabra computational framework, including the deep-layer temporal context mechanism, the combination of BCM-like Hebbian learning along with error-driven learning, and the effects of using momentum in the learning rule. Figure 17 shows that each of these factors plays an important role in contributing to the overall performance of the intact network. For the Hebbian and momentum factors, both of these produced more "dead" units (the flip-side of the hog

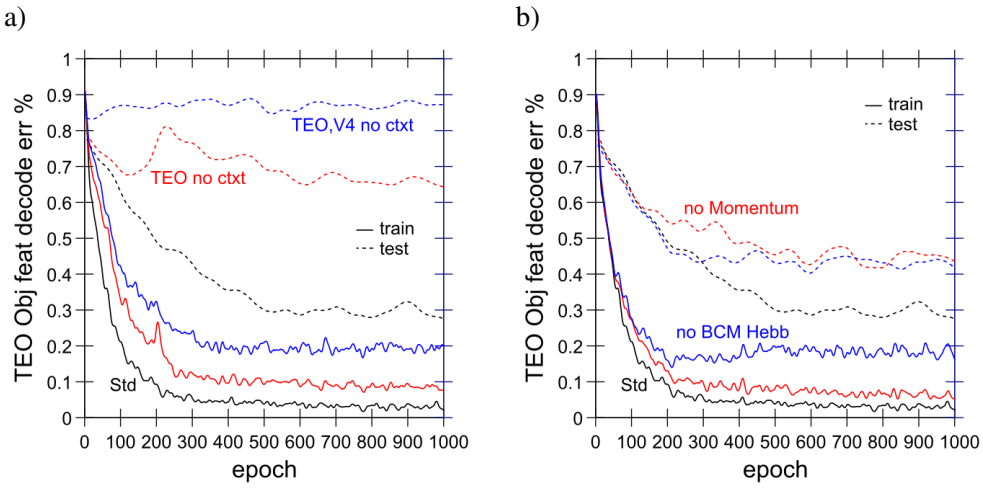

Figure 17: a) Effects of removing the deep-layer context inputs into TEO and V4 + TEO together — this has a major impact on ability to decode object features from TEO, particularly in the case of the novel testing items. b) Effects of not using momentum or BCM-like Hebbian learning.

units mentioned above — these are easier to quantify), particularly in the higher layers, with hebbian being particularly important for TEO while momentum was more important for V4.

Summary

The above results, which represent a small subset of the extensive explorations we performed over the development of the final model (1,160 different model runs, requiring over 45 CPU-*years* of computation on our 576 CPU cluster), together support a consistent overall picture of how the model learns over time. The three different pathways of the model, *Where*, *What*, and *What* * *Where*, interact in important ways to enable the joint goals of highly accurate prediction generation, and the development of invariant, systematic object representations in the ventral *What* pathway. This latter outcome depends on the other sources of prediction error being managed by other areas, and represents an important new way of understanding how a purely self-organizing learning system can develop these essential high-level abstract representations. In other words, this is a case where "it takes the whole network to raise a model" — the entire predictive learning problem must be solved with a complete, interacting network, and cannot be solved piece-wise. Furthermore, the entire network must be interacting bidirectionally, with top-down excitatory connections playing a critical role in shaping the overall learning process in lower layers, which then feed back up into the higher layers, etc. Thus, this model represents a truly *emergent* system.

Results: Accounting for Empirical Data

In this section, we apply our model to a set of important empirical phenomena that directly relate to predictive learning, starting with the case of predictive remapping, which is perhaps the most iconic example of a predictive phenomenon in the brain. We then simulate key data from monkey electrophysiology showing top-down effects emerging after roughly one alpha cycle, shaping lower-level representations according to higher-level interpretations of the overall scene. Finally, we simulate data that has been interpreted as supporting an alternative explicit-error-coding framework for generative models, showing that it emerges naturally from our model. Although these are but a small subset of the possible data within the scope of such a comprehensive model, they address some of the most important and relevant data.

Predictive remapping

The remarkable phenomenon of predictive remapping, where neurons in the visual stream appear to remap their spatial receptive field in anticipation of the effects of a saccade (Duhamel et al., 1992; Colby et al., 1997; Gottlieb et al., 1998; Nakamura & Colby, 2002; Neupane et al., 2016), is the exactly what one would expect if the brain is performing predictive learning. And indeed, our model was designed specifically to capture this effect, using saccades as one of the major sources of spatial prediction that the model needs to learn (the other being intrinsic motion of the object itself). Predictive remapping was initially described in area LIP (Duhamel et al., 1992), but it has also been found as low as V2 in the early visual stream, but, interestingly, not in V1 (Nakamura & Colby, 2002). In LIP, around the time of the saccade, neurons fire for stimuli that will appear in the new retinotopically-defined receptive field location, in anticipation of the effects of the saccade (Figure 18a).

Figure 18b shows the activity profiles of characteristic units in our model from LIP and V2 layers, providing a clear match to the observed data. Importantly, our model predicts that the remapping starts in LIP, which has direct input from the relevant eye movement signals, and this then drives top-down updating of activations in lower layers (V3, V2). Figure 19 shows this same trial in terms of full network activation patterns. This is consistent with the theoretical frameworks of Cavanagh et al. (2010) and Wurtz (2008), who strongly emphasize that this remapping must occur in these higher layers first, and then drive a top-down attentional signal to lower layers. It is simply not possible for lower layers to remap across the relevant visual angle of saccades, which can be quite far, and would require massive interconnectivity in these lower layers. It makes more sense for a compact, high-level spatial layer like LIP to do the essential spatial remapping, and then send the result down to lower layers. Critically, our model predicts that this top-down remapping

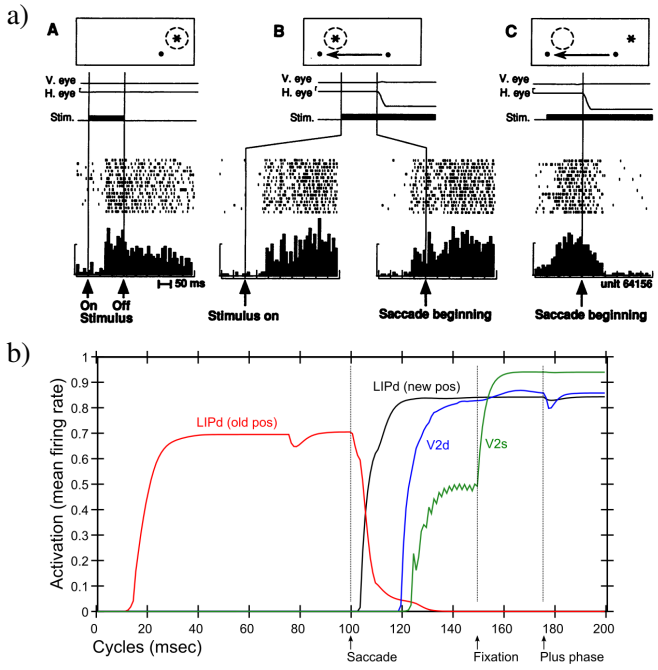

Figure 18: a) Original remapping data in LIP from Duhamel et al (1992). A) shows stimulus (star) response within receptive field (dashed circle) relative to fixation dot (upper right of fixation). B) Just prior to monkey making a saccade to new fixation (moving left), stimulus is turned on in receptive 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. b) 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 msec 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.

largely stops at V2, because that is the first layer that is driven by the predictive signals from the pulvinar — V1 is largely driven by LGN thalamus, and does not engage in this same kind of predictive learning process. This is consistent with available data (Nakamura & Colby, 2002), which also supports our prediction that V2 remapping is weaker and slower than that in LIP.

Our model makes some testable predictions about the relationship between saccades and the alpha cycle. For example, depth-electrode recording in LIP should be able to distinguish between a predictive representation emerging in the deep layers, strongly synchronized with the alpha cycle, and a more fluid superficial-layer representation reflecting current attentional foci, which is then updated via the predictive signals from the deep layers around the time of a saccade. We also predict that the pulvinar plays a critical role in broadcasting the predicted saccade outcome information to superficial LIP and other areas (along with LIP deep-layer top-down projections). Indeed, very recent data appears strongly consistent with these

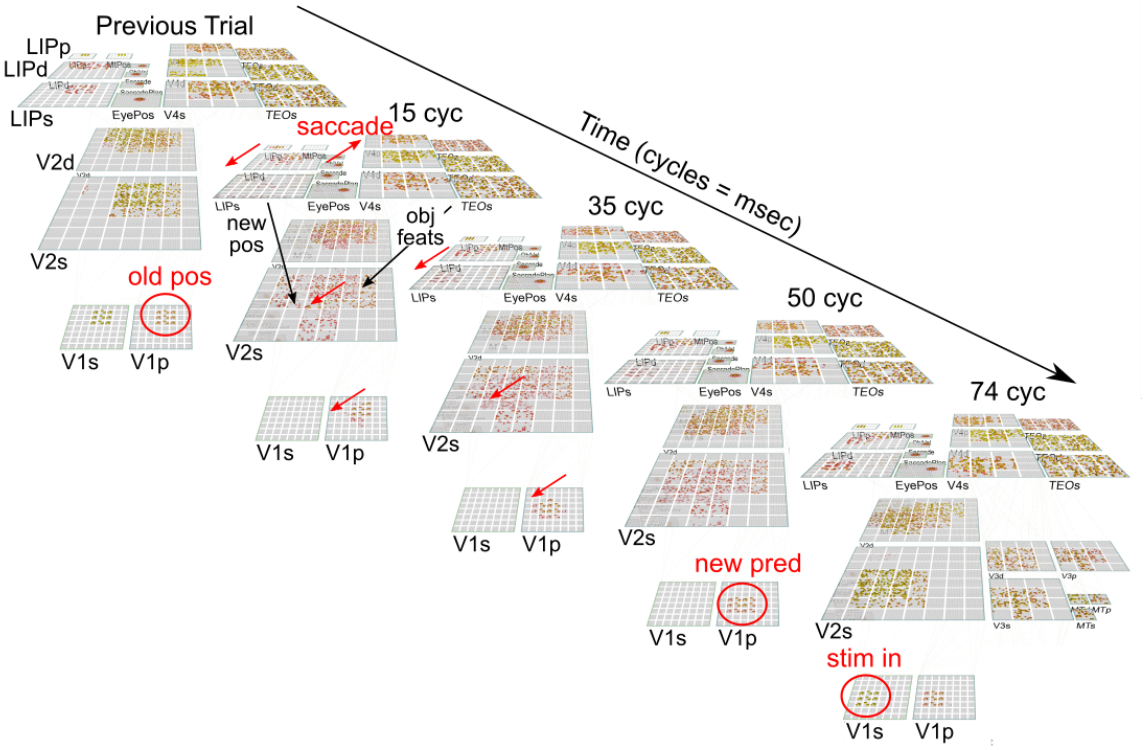

Figure 19: Predictive remapping in the entire model, from the prior trial state through the minus phase (75 cycles) of the post-saccade fixation trial. Though hard to see, the LIPd deep-layer activation state moves first within the first 15 cycles, which then drives LIPp (which then updates LIPs as well), and sends top-down input to V3 (not shown) and V2, which ultimately consolidates on a new predicted V1p state by 50 cycles. On cycle 74 (75th cycle), the new sensory input appears, matching the prediction. To make this accurate prediction, these lower layers receive top-down input from TEO providing a representation of object features, and these streams are combined (with considerable help from the V3/MT *What * Where* integration pathways) to drive an accurate prediction on the pulvinar (V1p) about what the visual input will look like when it arrives, after the saccade fixation.

predictions, showing a strong alpha-frequency coherence between the current and predicted receptive fields in V4, which they speculate to be driven by top-down and pulvinar-driven alpha dynamics (Neupane, Guitton, & Pack, 2017). This appears to be a very strong confirmation of a major prediction from our model.

Top-down Activation of V1 from Higher-Levels

There have been a number of important demonstrations that neurons in lower visual areas (V1, V2) reflect higher-level interpretations of a visual display, with this top-down signal emerging typically after around 100 msec (Supèr, Spekreijse, & Lamme, 2001; Fahrenfort, Scholte, & Lamme, 2008; Lee & Nguyen, 2001; Lee, Yang, Romero, & Mumford, 2002) (Figure 20). Importantly, these effects depend on the animal being awake, and on having indicated that the higher-level percept was actually formed (Supèr et al., 2001), and other factors such as context that shape the nature of the high-level interpretation (Lee & Mumford, 2003). Given the importance of top-down activation from higher layers in our model, we tested for the presence

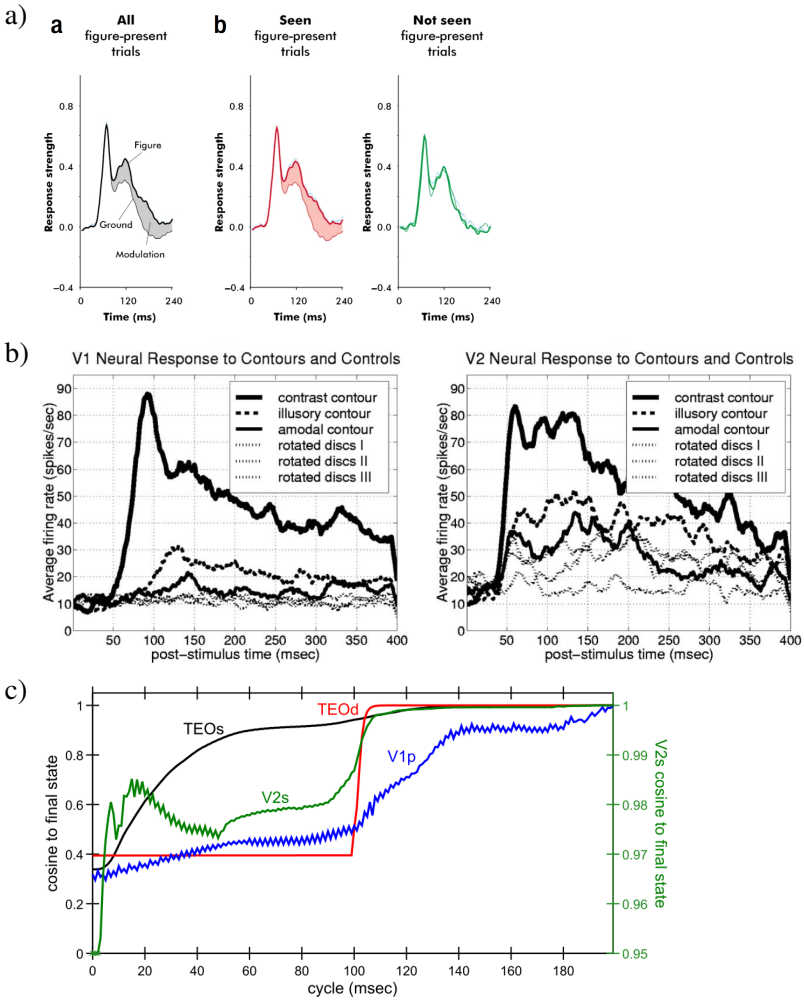

Figure 20: Top-down effects on lower-level neural firing. a) Top-down modulation of V1 firing as a function of a texture-defined figure/ground stimulus, emerging after 100 msec (one alpha trial) in monkeys, specifically as a function of whether the monkey makes a behavioral response indicating that the figure was seen (Super et al, 2001, reprinted with permission). b) Emergence of V1, V2 neural firing to illusory and amodal contours, suggesting earlier V2 responding driving top-down V1 responses that emerge after 100 msec (Lee et al., 2002, reprinted with permission). c) Top-down driven activation in V1 and V2 of our model, using partially-occluded stimuli, showing the cosine of the current activity pattern on a layer in comparison to the final activation state at the end of the 200 msec window. TEOs (superficial neurons) converges on its final state the most quickly, and drives top-down updating of V2s and V1p (pulvinar) representations, which are then more strongly driven when the TEO deep-layer (TEOd) updates after one alpha cycle. The final V1p state reflects a largely accurate prediction of the complete object features (see supplemental information for a video of actual network states) — the remaining change at the very end reflects plus-phase signal driving back to partial input, which does not perturb higher layers. Note that V2s is plotted on a separate scale (shown at right) because it is a very large activation pattern that doesn't change as much as the others.

of similar such effects. Because of the simplicity of our visual environment, we could not directly replicate the existing experiments (which involve 2D-cues for depth perception), but instead used a simple proxy, where the object inputs were partially obscured (11% of active features turned off), such that higher-level representations were needed to complete the original full pattern.

As Figure 20 shows, our model shows the same kind of top-down effects in lower layers as have been

observed in monkeys (and in our prior bidirectional object-recognition model; O'Reilly et al., 2013). The consistent observation that these top-down effects emerge just after 100 msec is consistent with the importance of deep-layer updating at the alpha rhythm (and the relative importance of deep-layer projections for top-down activation), which is an essential property of our model.

Activation Differences between Predicted and Unpredicted Inputs

As we review more extensively in the General Discussion section, there is an important difference between our model and many other types of generative models, which postulate the presence of neurons that explicitly code for the mismatch error between the top-down generated model and the bottom-up sensory input (Mumford, 1992; Rao & Ballard, 1999; Kawato, Hayakawa, & Inui, 1993; Friston, 2005). Under these frameworks, top-down pathways have a net inhibitory effect on lower-level neurons, subtracting away predicted aspects of the signal. This is the opposite of the excitatory top-down effects just shown above, where top-down excitation can fill in missing elements and shape the representation to accentuate lower-level elements that are consistent with the higher-level interpretation of a scene.

Nevertheless, there are various sources of evidence that have been seen to support these explicit error-coding models, principally the finding of relatively less activation for predicted versus unexpected outcomes (e.g., Summerfield, Trittschuh, Monti, Mesulam, & Egner, 2008; Todorovic, van Ede, Maris, & de Lange, 2011; Meyer & Olson, 2011; Bastos, Usrey, Adams, Mangun, Fries, & Friston, 2012) (sometimes the opposite result is found; Anderson & Sheinberg, 2008). However, there are a number of alternative mechanisms that can account for this same pattern, and various attempts to systematically evaluate the available evidence have been inconclusive and somewhat mutually contradictory (Kok & de Lange, 2015; Kok, Jehee, & de Lange, 2012; Summerfield & Egner, 2009; Lee & Mumford, 2003). None of these reviews concludes that there is any solid *direct* evidence for explicit error coding, including the most recent one (Kok & de Lange, 2015), but they nevertheless reach different overall conclusions based on the overall body of indirect evidence, much of which comes from human neuroimaging studies and is subject to various forms of alternative explanations.

Here, we explore the extent to which our model, which definitely lacks any such explicit error coding neurons, can account for some of the observed patterns of data. First, to review some of the major alternative explanations, there are well-established temporal dynamics of neural firing that naturally cause neurons to reduce their firing level over time, lasting for different time scales. As is evident in just about every electrophysiological recording in neocortex (e.g., Figure 20a,b) neurons typically exhibit a large initial transient burst of activation, followed by a slower decrease in firing rate over the next several hundred

milliseconds. Some of the initial burst may be due to delay in onset of inhibitory feedback mechanisms, and there are also well-documented rapid-onset, transient spike frequency adaptation mechanisms that are essential for accurately capturing pyramidal cell firing patterns (Brette & Gerstner, 2005; Gerstner & Naud, 2009). Lasting slightly longer are synaptic depression effects (Markram & Tsodyks, 1996; Abbott, Varela, Sen, & Nelson, 1997; Hennig, 2013) which can account for several important aspects of neural adaptation (Müller, Metha, Krauskopf, & Lennie, 1999). At a yet longer time-scale, fast synaptic plasticity interacting with inhibitory dynamics can account for an overall *sharpening* phenomenon across distributed neural representations, where the tuning of active neurons becomes narrower and more selective, while weak, broadly-tuned neurons drop out, resulting in an overall net reduction in neural activation (Desimone, 1996; Wiggs & Martin, 1998; Norman & O'Reilly, 2003). This sharpening dynamic is considered likely to underlie many aspects of the *repetition suppression* effect widely-observed in human neuroimaging studies (Grill-Spector, Henson, & Martin, 2006), and many of the phenomena typically offered in support of explicit error-coding are also consistent with a sharpening-based account (Kok et al., 2012; Lee & Mumford, 2003).

One clear way in which the above mechanisms could produce a seeming inhibition of inputs that are consistent with a prediction, is if the prediction process drives top-down activation of relevant neural representations *in advance of stimulus input*, such that these representations are *already* adapted / depressed / sharpened by the time the stimulus arrives. It is unclear why this kind of effect would *not* arise, and it should account for all of the same prediction-dependent phenomena as the explicit error-coding account. However, our current model does not have any of the above basic adaptation, synaptic depression, or fast synaptic plasticity mechanisms turned on (although all of them are available in our simulator) — we will more systematically investigate this type of explanation in future work.

It is also possible that the reduced level of activation for predicted outcomes is simply due to the transition whereby neurons activated as a consequence of prediction are decaying while neurons activated by the actual unpredicted input are firing. The long time scale of human neuroimaging would show an overall increase in activation. There is a larger "smear" of neural activation over time in the unpredictable case compared to a case where a single stable representation is active over time (i.e., the predicted outcome actually occurs). Any additional suppression of these stable representations over time would only accentuate the magnitude of the difference between unpredicted and predicted, as it would differentially affect the stable predicted representations.

Figure 21b shows this transition effect for layer V2s comparing the more predictable pre-saccade trial (2nd trial) with the post-saccade trial (3rd trial), which is less predictable due to residual difficulty in fully predicting the outcome of the saccade. Because the V2 layer is highly retinotopically organized, it shows

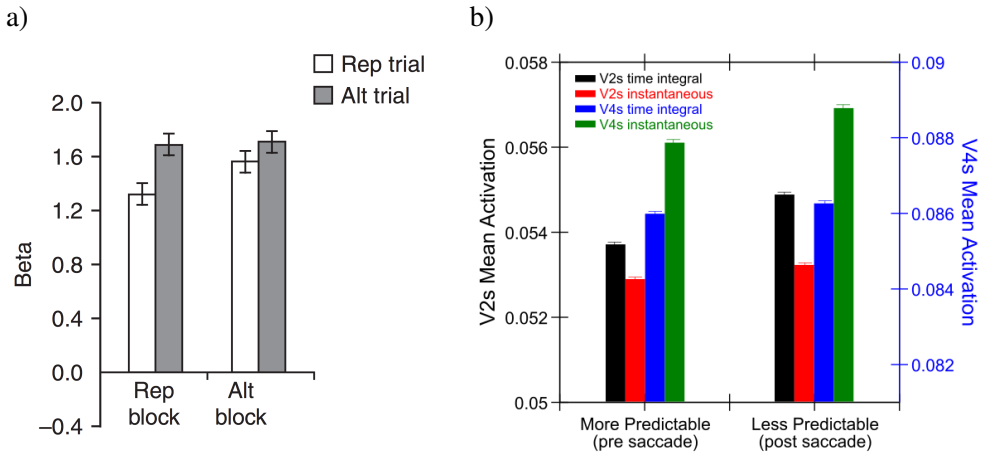

Figure 21: Activation reductions for more vs. less predictable trials a) Data from Summerfield et al. (2008) fMRI study, comparing a block-wise manipulation of probability of repetition (75% for Rep block, 25% for Alt block). Repetition suppression is enhanced when repetitions are more expected (Rep block). b) Results from our model, on the trial before saccade (2nd trial) which is more predictable based on first trial inputs compared to the immediate post-saccade trial (2nd trial), which is less predictable due to the residual difficulty in fully predicting saccade outcome. The V2s layer shows a significant increase in time-averaged activation across the trial for the less predictable case (black bars), even though this is not seen in instantaneous activations (red). Higher up in V4 we see the reverse pattern, where instantaneous activation (green) is higher for the less predictable case, but the time-average does not differ — this is, we believe, because V4 activation is not as topographically specific, but it does perform its own time integral over V2.

increased time-averaged activation when predictions do not quite align with the new inputs compared to the relatively more stable time-averaged activation in the second trial. This is true even though the instantaneous activation (recorded at the end of the trial) is essentially the same. The same patterns were seen in the V2s (superficial) and V2d (deep) layers (not shown). In contrast, at the higher, less topographically-organized V4 layer, there is much less difference in time-averaged activation across the two trials. However, by the end of the third trial, the instantaneous activation is somewhat higher, presumably because V4 is itself integrating over the V2 layer. This effect is not present in the next higher (TEO) layer (not shown).

For comparison, Figure 21a shows fMRI data from Summerfield et al. (2008) that has been interpreted as supporting the existence of explicit error coding neurons. They compared cases where stimulus repetitions (faces) were more or less predictable, and found more of a repetition suppression effect in the more predictable case. In the context of our model, we would say that people formed stronger predictions of the face repeating in the 75% repetition block and there would be little difference in neurons activated by prediction and those activated by the actual input. Interestingly, there was no effect of reducing activation in the alternating case when alternations were 75% of trials, even though the alternation was more "predictable" — it is impossible to form a concrete prediction for the alternation case, so whatever face does show up there is a surprise from a visual prediction standpoint, and results in equivalent amounts of activation from predicted

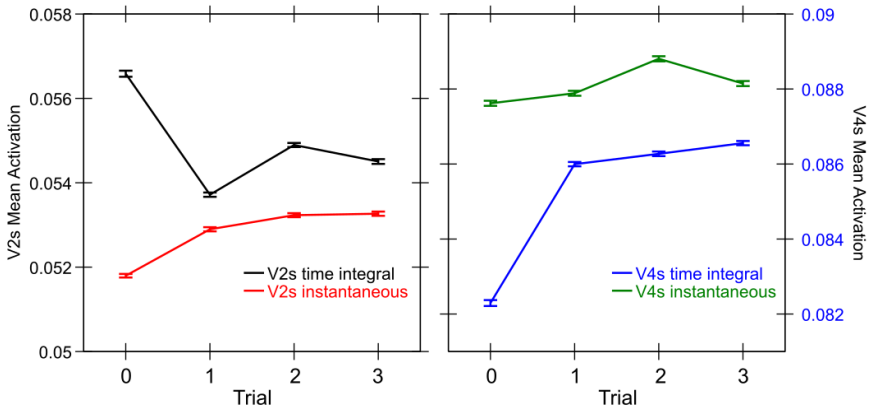

Figure 22: Time-integral and instantaneous activation across all 4 trials for V2s and V4s layers (data in previous graph comes from Trials 1,2). Overall, trial 3 should be similarly predictable as trial 1, and the activations are consistent with this. Trial 0 is highly unpredictable, and shows even higher levels of churn in V2s time-integral, while overall having lower instantaneous activity. V4 is ramping up to its final activation during this trial and thus the time integral is lower.

to actual in all cases.

Finally, Figure 22 shows all four trials to give a fuller picture of the activation dynamics, and further evidence that the activation increases selectively in the more unpredictable post-saccade trial (Trial 2, the 3rd trial) compared to both of the surrounding more-predictable trials. Also, we separated the data according to trials where the object was correctly decoded from TEO from those where it was not (all of the above data are from correct trials). The error trials overall showed similar patterns of activation, but, interestingly, exhibited a consistent and sizeable reduction activation overall across all the trials (a difference of about .02 in V2s). This is consistent with the idea that overall network coherence and representational strength is important for accurate performance, as is often found in electrophysiological correlates of behavior.

In summary, these analyses demonstrate a novel origin for observed relative reductions in (time-averaged) activation for more predictable vs. more unpredictable trials. We anticipate that adding the various forms of repetition suppression mechanisms mentioned above will only increase the strength and robustness of these basic effects, and then it would be appropriate to make a number of more strongly testable predictions from the model. One clear prediction from the model is that higher brain areas can integrate over "churn" present in lower areas, to produce in instantaneous activation that is only present in time-averaged activation at the lower level. While any small set of data points may be consistent with a variety of models, comparing error vs. correct performance across a variety of trial types, layers, and neural measures should prove strongly constraining.

General Discussion

We have presented a comprehensive model of the visual system that demonstrates how predictive learning within a generative framework leads to high-level invariant object representations without any external training signal. The model follows known biology and accounts for data across many levels of analysis, from low-level synaptic plasticity to systems-level organization and connectivity of the areas and pathways of the visual system, including the development of these pathways. The pulvinar nucleus of the thalamus plays a central role as a kind of projection screen, upon which the different visual areas across levels of abstraction collaboratively project their predictions for what the visual input will look like when the next alpha-frequency (100 msec) 5IB driver inputs provide their ground truth plus-phase training signal. The pulvinar broadcasts back out to all the areas that contribute to it, enabling neurons everywhere to learn based on the temporal difference between the minus-phase prediction and plus-phase target. Synaptic plasticity mechanisms capable of using this temporal difference were derived directly from a biophysically detailed model of spike-timing dependent plasticity (Urakubo et al., 2008). Computationally, the direct and indirect propagation of this prediction error signal produces powerful error-backpropagation learning, capable of shaping deep hierarchies of representations to minimize the prediction error.

The collective prediction error signal from the pulvinar is partitioned into three separable components by three different visual pathways: *Where*, *What*, and *What * Where* integration, through a combination of developmental sequencing and emergent dynamics of learning shaped by specific patterns of interconnectivity. This allows compact, high-level, abstract representations at the top of each of these pathways to drive low-level predictions, which is essential for successful predictive learning, as the lower-level areas are too retinotopically diffuse to provide effective predictive representations over time. The particular developmental and connectivity constraints that emerge from these principles, along with the results of extensive experimentation in our model, align remarkably well with available data on the primate visual system.

To summarize, here are some of the major, well-established biological properties that are central to our model:

- A strong synchronized, low-frequency modulation of cortex (at the alpha frequency), specifically in the deep layers and thalamus, as opposed to superficial layers.
- Nature of deep-layer connectivity to pulvinar, specifically having *both* a numerous, weaker, plastic pathway (for generating a prediction) and a sparse, strong, fixed pathway (for providing a *ground truth* target).

- Synchronization of this strong pathway input with the alpha cycle.
- Broad connectivity of pulvinar with different visual pathways (afferent and efferent).
- Lack of direct bottom-up superficial projections into the deep layers, but presence of these projections top-down.
- Bidirectional (top-down and bottom-up) connectivity between superficial layers of connecting areas.
- Early development of the *Where* (MT, LIP) pathway.
- Organization into three separable (yet highly interconnected) visual pathways, particularly a third putative *What* * *Where* integration pathway.

While there are various other theoretical interpretations of each of these different phenomena, we are not aware of another framework that ties together all these different elements under an overarching computational model. Furthermore, we argue that our model provides a theoretical continuity between levels of analysis that have previously not been well-aligned. For example, biologists tend to think that the brain learns using Hebbian learning mechanisms, but computationally these are very limited, and computer scientists have overwhelmingly embraced error-driven backpropagation models instead. However, error-backpropagation is widely regarded as biologically implausible for a variety of reasons (e.g., Crick, 1989), not all of which are resolved by local, activation-based versions (O'Reilly, 1996; Movellan, 1990; Xie & Seung, 2003; Scellier & Bengio, 2017). One of the most important unresolved such issue is the question of where the error signals actually come from to drive backpropagation — current models rely extensively on large human-labeled datasets. Thus, the ability of our model to provide a biologically-sound framework for powerful error-backpropagation learning using only raw sensory streams, through the principle of predictive learning, establishes a clear theoretical continuity between levels.

As such, we offer it as a possible answer to the longstanding mystery of how the neocortex develops and learns over the first several months of life to produce the foundations of all our high-level cognitive abilities. In particular, the finding that this purely self-organizing predictive learning process, in combination with all the systems-level structure in which it is embedded, can form systematic invariant object representations characteristic of those found in TEO and other IT areas, provides a foundation for subsequent categorization. We are excited to extend our model with auditory pathways, to understand how combined multi-modal predictive learning across vision and audition interact in this next level of cognitive learning (which also likely shapes the nature of visual learning in important ways not captured in the present model). Preliminary

work in this direction using earlier versions of our predictive learning framework suggests that the auditory pathway is highly amenable to predictive learning approaches in general, due to the intrinsically temporal nature of auditory signals, consistent with the success of predictive learning frameworks in linguistic datasets (Elman, 1990, 1991; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013).

In the remainder of the discussion, we compare this framework with other related frameworks, consider some broader implications of our approach, and then highlight a few of the many central testable predictions from our model, followed by a further discussion of a number of unresolved questions for future research.

Comparison with other Frameworks

Generative Models

Our framework fits within the broader context of *generative models* in psychology and neuroscience, which embody the principle of *recognition by synthesis*, which goes back at least to Helmholtz in 1867 von Helmholtz (2013). This idea was advanced by a number of different researchers in various ways in the 1990's as a possible way of understanding neural function (Mumford, 1992; Kawato et al., 1993; Ullman, 1995; Dayan, Hinton, Neal, & Zemel, 1995; Rao & Ballard, 1999), with Carpenter and Grossberg (1987) having a somewhat different but related earlier framework. Common to most of these frameworks is the notion of a hierarchy of areas stacked upon each other, with higher layers having more abstract, compact internal models of the environment, and some kind of interplay between a feedforward pathway of sensory information flowing up this hierarchy, and a feedback pathway driving top-down signals based on internal generative models.

Most of these models (Mumford, 1992; Kawato et al., 1993; Dayan et al., 1995; Rao & Ballard, 1999) adopt an *explicit error-coding* framework, where certain neurons explicitly subtract the top-down model-based signals from the bottom-up sensory-driven signals, to represent the mismatch between the two (while another population represents the accumulated top-down prediction itself). This error signal is typically fed forward to higher layers, which then use it to adjust their current model parameters to better fit with the sensory inputs, in an iterative process. Somewhat confusingly, these error signals are sometimes referred to as "prediction errors" but this sense of the word prediction does not typically include the critical "about the future" aspect — they are usually just static "predictions" of the current sensory inputs, from the generative model (a more appropriate term would be *generative errors* or something to that effect). Mumford (1992) hypothesized that the neocortical superficial layer neurons encode this error signal and project it feedforward, while the deep layers transmit the model-based predictions top-down — this same idea was also advocated by others (Rao & Ballard, 1999; Kawato et al., 1993). Carpenter and Grossberg (1987)

adopted a more discretized, localist version of this process, where a single upper-layer neuron is activated (representing the internal model), and the degree of mismatch between its top-down weights and the current stimulus is used, with a sensitivity threshold, to determine whether to keep that neuron active, or select a new one to encode the current input stimulus.

The hierarchical generative model idea was embraced and further developed with the subsequent popularity of the Bayesian framework, where it has a direct and clear relationship to the key Bayesian twist (e.g., Lee & Mumford, 2003; Friston, 2005; Yuille & Kersten, 2006; Friston, 2008, 2010; Lee, 2015). This Bayesian twist turns a question about how likely various hypotheses (models) are given some observed data, into the question of how likely the *data* is given various hypotheses (i.e, the generative model). The latter form is typically much easier to compute, and inference (going from the data to the model) can be performed by adapting the model to more closely generate the observed data, as proposed in these early neural models. In machine learning and statistics, a generative model has a more formal definition in terms of capturing the full probability distribution of the data, and a well-defined formal probabilistic structure such that it truly can generate plausible data *de novo*. In contrast, our use of the term as a model of brain function is much looser, including all such models that include any aspect of a generative process, such as neural network auto-encoders.

In contrast to the above models, the counter-streams model of Ullman (1995) holds that the feedforward and feedback pathways are collaborative and amplify areas of congruence or match between top-down and bottom-up pathways. This is more in the spirit of the bidirectional constraint satisfaction framework that is a foundation of our approach, based on earlier frameworks developed in the 1980's (Hopfield, 1982, 1984; Ackley et al., 1985; Rumelhart & McClelland, 1982). In this overall framework, the activation states for both the superficial and deep layers *always represent the best guess internal representation of the sensory inputs*, not a difference or error signal. This allows both top-down and bottom-up signals to converge on shaping these internal representation states in a collaborative way (i.e., bidirectional constraint satisfaction), instead of positing a fundamentally subtractive or contrastive relationship between the bottom-up and top-down pathways. As we have demonstrated, this excitatory, collaborative influence of top-down inputs is critical for allowing high-level abstract representations to shape accurate low-level predictions in our model.

In sum, there is a fundamental division between frameworks based on the principle that bottom-up and top-down streams have a net subtractive, mismatch-coding relationship, versus those based on a more collaborative, match-amplification dynamic between the two streams (the deep layers in the mismatch-coding generative models do exhibit this match-amplification property, so the contrast here is focused specifically on the hypothesized superficial error-coding neurons). Computationally, there may be a critical difference

between these approaches in terms of how effectively they converge on an interpretation of the current sensory input. Intuitively, this difference corresponds to the difference between the "Yes, and.." approach to collaborative problem solving, versus the "No, but.." approach, as highlighted in a popular book by comedy writers (Leonard & Yorton, 2015). The collaborative, positive approach brings *all* of the available constraints (top-down and bottom-up) to bear on rapidly converging on a reasonable interpretation. In contrast, the error-based generative models are dominated by critical negative input from the top-down pathway, which is great for eliminating bad interpretations but not for collaboratively finding good ones. Also, the strictly hierarchical nature of most generative models, where each layer serves exclusively as the model for the layer below it, may make the inference process more difficult. In contrast, all of the different levels of abstraction in our model collaborate together to produce a single integrated prediction, projected onto the pulvinar "silver screen of the Cartesian theater." The broad projections from pulvinar back to cortex then share this developing prediction with all the relevant contributing layers, helping to coordinate all levels together simultaneously, instead of each working separately on their own relatively isolated problem.

Instead of using error signals during the online inference process, we think they are more effectively used to guide the learning process, which takes place over a much longer time period, and only needs to converge once. Here, the stochastic gradient descent process embodied by the error backpropagation algorithm has consistently proven its value as a way of optimizing learning in deep hierarchical networks.

Biologically, we reviewed above the evidence bearing on whether superficial layer neurons in the neocortex encode prediction errors, and showed that our model can account for the key finding of reduced activation for predicted relative to unpredicted events. This and other alternative accounts of the main indirect evidence for explicit error-coding neurons, together with the notable lack of any solid direct evidence for this central hypothesis of most generative model frameworks, should be sufficient to render such a framework biologically implausible at best. More generally, there are so many detailed electrophysiological recordings of neurons throughout the cortex showing that neural firing positively encodes representations of the current environment, that it seems rather unlikely there could be a large population of explicit error-coding neurons lurking in there somewhere. Furthermore, the idea that feedback projections are inhibitory is at odds with the basic anatomy, where all long-range connections in the neocortex are excitatory (Johnson & Burkhalter, 1997; Shao & Burkhalter, 1996), and the excitatory nature of these top-down connections is compatible with the well-supported biased-competition model (Desimone & Duncan, 1995; Miller & Cohen, 2001). Although there are ways of reshuffling connections to make biased-competition and generative models more mathematically consistent (Spratling, 2008), this approach still retains the requirement of inhibitory top-down connections (biased competition is made to be more like a generative model, where lateral pooled

inhibition is replaced with top-down inhibition, and also activations and synapses that can be either positive or negative), which Spratling (2008) acknowledges are biologically implausible.

In summary, although our framework shares the overall generative model goal, it achieves this goal in a fundamentally different way from most generative models, which we argue has both computational and biological plausibility advantages. Furthermore, our model is distinct in being architecturally founded on making true predictions about the future, instead of just re-generating the current sensory inputs. Despite these differences, it is likely that many of these theorists would recognize our model as fitting well within their broader vision for how neocortex works.

Deep Auto-encoder Neural Networks

The Restricted Boltzmann Machine (RBM) framework (Hinton, 2002; Hinton & Salakhutdinov, 2006; Hinton, 2007) represented a critical bridge between the Bayesian generative model framework, and the now-dominant resurgence of neural network models. The RBM was derived from a mathematically well-characterized generative-model framework, but required a final training phase using error backpropagation. Eventually, it became apparent that the initial RBM training could be skipped entirely, with the development of various important tricks for making deep (i.e., having many hidden layers) models converge effectively (Ciresan, Meier, Gambardella, & Schmidhuber, 2010; Ciresan, Meier, & Schmidhuber, 2012; Krizhevsky, Sutskever, & Hinton, 2012; Bengio, Courville, & Vincent, 2013a; LeCun, Bengio, & Hinton, 2015). One of the most important such tricks is the use of weight sharing among topographically organized groups of units in lower layers, which mathematically is the same as *convolution* by a filter defined by this set of shared weights (LeCun, Boser, Denker, Henderson, Howard, Hubbard, & Jackel, 1990; LeCun et al., 2015).

Most of the deep neural networks (i.e., *deep nets*) are trained to produce localist category labels for bitmap images, and do not include generative-model aspects. Nevertheless, these models appear to capture some important properties of the ventral *What* pathway (e.g., Majaj et al., 2015), building on insights from earlier more neuroscience-inspired frameworks (Riesenhuber & Poggio, 1999). However, they require vast amounts of hand-labeled image data, and are thus not plausible models of the largely self-organizing nature of human visual learning. Indeed, we argue that these models are somewhat like powerful 3D printers, that instead print brain circuits mimicking those in the human brain. Their performance is proportional to the sample size of human behavior available (e.g., number of samples of human object categorization applied to a wide range of images), which is analogous to how fine-grained the scan of an object is for a 3D printer — the finer the scan, the more accurate the reproduction. Because the mapping function from image to object label present in human brains is very high-dimensional, a very large number of samples is needed to

reproduce it accurately. To continue the analogy, a deep convolutional neural net also constitutes a good raw material to "render" in, as it starts out with structural biases etc that match those of the visual system. And, several tricks that improve performance are also biologically-supported properties such as winner-take-all learning and pressure to develop sparse representations, which are also included in our Leabra framework. By contrast, our model represents an attempt to reconstruct the complex interactive dynamics that shape the human visual system based on raw visual input, without relying on any direct sampling of the mature system.

There has also been some renewed focus on deep versions of auto-encoder models, which are the neural network equivalent of a generative model (Bengio, Yao, Alain, & Vincent, 2013b; Valpola, 2014; Rasmus et al., 2015; Le et al., 2012). Many of these models adopt a denoising training strategy to prevent the model from just learning a degenerate "copy the input" strategy (Bengio et al., 2013b), and include a strongly hierarchical outside-in training strategy in the form of a *ladder* network (Valpola, 2014; Rasmus et al., 2015). Very recently, this auto-encoder paradigm has been extended into a true predictive learning framework like that in the present model (Lotter, Kreiman, & Cox, 2016). This model is trained in a purely unsupervised manner on movies, predicting the next frame, which is effectively what we are doing. The model learns to generate realistic-looking images and achieves overall good predictive error scores. The analysis of the internal learned representations focused on lower-level visual parameters such as camera pan and roll, and there did not appear to be any invariant object representations that self-organized. The model was also trained to decode faces using subsequent supervisory training, with similar overall results to comparable auto-encoders.

Thus, there are considerable similarities at a broad level between these models and our framework, but overall these models are more closely aligned with traditional Bayesian generative models than our framework. For example, they adopt a strict hierarchical structure to the layers, with each higher layer attempting to encode the layer below it, instead of the multi-pathway, collaborative-across-levels approach characteristic of our model. Furthermore, they do not typically include any bidirectional constraint satisfaction processing, so the inference process is strictly feedforward. Finally, these models are not used in a purely self-organizing manner — the final step is generally to train on standard human-labeled supervised datasets, and the key measure of interest is the extent to which the auto-encoder pretraining reduces the amount of supervised training required to achieve a given level of performance.

Biologically, there has been a long history of skepticism about the biological plausibility of error-driven backpropagation learning (e.g., Crick, 1989). As noted earlier, we have long argued that these issues can be overcome through the use of bidirectional excitatory connectivity and temporal-difference based synaptic

plasticity, which closely approximate error backpropagation (O'Reilly, 1996) (see also Movellan, 1990; Xie & Seung, 2003; Scellier & Bengio, 2017). Furthermore, we have shown how models using these learning mechanisms can learn like these other deep neural networks, while also exhibiting important bidirectional dynamics (O'Reilly et al., 2013; Wyatte, Herd, Mingus, & O'Reilly, 2012b; Wyatte, Curran, & O'Reilly, 2012a).

Forward Models

A major, well-established application of predictive learning is for *forward models* that predict the outcome of actions (Kawato et al., 1987; Jordan & Rumelhart, 1992; Miall & Wolpert, 1996). The LIP predictive remapping from saccades is really a form of forward model (predicting the next sensory state that follows from the motor action of moving the eyes), and our model advances the idea that every area of cortex has a deep-layer forward model associated with it. Besides driving the self-organization of the entire visual system, one might ask what other potential benefits all these forward models might have? One popular idea is that they can be used to select actions that achieve desired outcomes, by effectively running them backward (Hommel, 2004; James, 1890; Pezzulo & Castelfranchi, 2009; Friston, 2010). Although this *ideomotor* principle is attractive, it is not clear if it is tractable for realistic motor actions (Herbort & Butz, 2012; Jordan & Rumelhart, 1992). We are particularly skeptical of prevalent models that hypothesize long sequences of chained predictions to generate action plans (Burgess & O'Keefe, 1997; Pastalkova, Itskov, Amarasingham, & Buzsáki, 2008; Lisman & Redish, 2009). Such chains are only as strong as their weakest links, and the working memory demands required to keep such a process going seem excessive, especially for rodents. Instead, we suggest that one-step predictions can be generated over many different time scales, and particularly in the prefrontal cortex, longer-time-scale predictions of outcomes are used to guide planful action (O'Reilly, Hazy, Mollick, Mackie, & Herd, 2014a; O'Reilly, Petrov, Cohen, Lebiere, Herd, & Kriete, 2014b; O'Reilly et al., 2016). Nevertheless, it is plausible that the same basic predictive learning mechanisms exploited in posterior cortex for fast-time-scale predictive learning could also be important for these longer-time-scale learning processes in frontal areas.

Due to the simple one-to-one retinotopic nature of saccade motor plans relative to the current visual input, this domain does not capture the more general challenges in motor learning. Therefore, we plan to explore the motor control implications of pervasive predictive learning in the context of the auditory pathway, including predicting the effects of speech output, to study the process of learning to imitate speech sounds, as has been explored using forward models (Guenther & Vladusich, 2012).

One major issue raised in this context is the relationship between the hypothesized forward models

learned in the cerebellum (Wolpert, Miall, & Kawato, 1998; Verduzco-Flores & O'Reilly, 2015; Shadmehr, 2017) relative to those in the neocortex. Although both systems may be learning predictive models, the cerebellum appears to be specialized for shorter, faster time scales of motor control (e.g., with around 10 msec resolution). Furthermore, differential effects of cerebellar lesions early vs. later in life suggest that the cerebellum serves to shape learning in the neocortex, which can then take on much of the learned functionality. The primary cortical output of the cerebellum goes to frontal and some parietal thalamic areas (Strick, Dum, & Fiez, 2009), so it may teach cortex by providing a plus-phase training signal, thereby plugging directly into the same learning system described here (similar to the superior colliculus inputs to the second pulvinar map as mentioned above; Shipp, 2003). We will investigate this possibility in future work.

Hawkins' Model

The importance of predictive learning and temporal context are central to the theory advanced by Jeff Hawkins (Hawkins & Blakeslee, 2004). This theoretical framework has been implemented in various ways, and mapped onto the neocortex (George & Hawkins, 2009). In one incarnation, the model is similar to the Bayesian generative models described above, and many of the same issues apply (e.g., this model predicts explicit error coding neurons, among a variety of other response types). Another more recent incarnation diverges from the Bayesian framework, and adopts various heuristic mechanisms for constructing temporal context representations and performing inference and learning. We think our model provides a computationally more powerful mechanism for learning how to use temporal context information, and learning in general, based on error-driven learning mechanisms. At the biological level, the two frameworks appear to make a number of distinctive predictions that could be explicitly tested, although enumerating these is beyond the scope of this paper.

Granger's Model

Another model which has a detailed mapping onto the thalamocortical circuitry was developed by Granger and colleagues (Rodriguez, Whitson, & Granger, 2004). The central idea behind this model is that there are multiple waves of sensory processing, and each is progressively differentiated from the previous ones, producing a temporally-extended sequence of increasingly elaborated categorical encodings (*iterative hierarchical clustering*). The framework also hypothesizes that temporal sequences are encoded via a chaining-based mechanism. In contrast with the DeepLeabra framework, there does not appear to be a predictive learning element to this theory, nor does it address the functional significance of the alpha frequency modulation of these circuits.

Other Frameworks for Cortical Oscillations

There have been a number of different computational functions ascribed to cortical oscillations and synchrony, which are not reflected in our model. Perhaps the most influential such idea is that different phases of cortical synchrony can support multiple interleaved *bindings* of separate features (e.g., Wang, Buhmann, & von der Malsburg, 1990; Gray, Engel, König, & Singer, 1992; Engel, König, Kreiter, Schillen, & Singer, 1992; Zemel, Williams, & Mozer, 1995; Hummel & Biederman, 1992). We have argued against such models in favor of coarse-coded distributed representations that naturally support binding without requiring an elaborate and brittle synchrony-based mechanism that ultimately requires decoding mechanisms that obviate most of the benefit of the binding in the first place (O'Reilly & Busby, 2002; O'Reilly et al., 2003; Cer & O'Reilly, 2006; O'Reilly et al., 2014b). The function of cortical oscillations in the current model serve instead to coordinate and organize the entire distributed network, which is generally widely accepted and uncontroversial. We have also developed models of the role of the theta rhythm in the hippocampus (Ketz, Morkonda, & O'Reilly, 2013), and the beta rhythm in the basal ganglia (BG) and prefrontal cortex (PFC) (Ketz, Jensen, & O'Reilly, 2015; O'Reilly et al., 2014b; Jilk, Lebiere, O'Reilly, & Anderson, 2008).

Briefly, we think that the hippocampal episodic memory system integrates over two alpha cycles in its theta frequency (5 hz, 200 msec) encoding and retrieval cycle, while the BG/PFC system operates at a faster cycle rate (beta = 20 hz, 50 msec) to allow more rapid behavioral responding and updating of working memory representations. Interestingly, the 50 msec time frame for BG function was independently established in the ACT-R model based on fitting behavioral data (Stocco, Lebiere, & Anderson, 2010; Anderson & Lebiere, 1998; Jilk et al., 2008). These functional roles contrast with the influential model of Lisman and colleagues, based on the numerical observation that 8 or so 40 hz gamma cycles can be embedded in one theta cycle, which seemed to correspond to the "magic number 7" working memory capacity constraint (Idiart & Lisman, 1995; Lisman & Jensen, 2013). However, outside of specialized phonological processing pathways, the pervasive representational capacity of any given brain area appears to be more like 2-4 (Cowan, 2001), and may have more to do with use of the two different hemispheres plus the ability to (barely) support at most two different distributed representations within a given area (Buschman, Siegel, Roy, & Miller, 2011).

Hinton's Joint View and Object Model

One of the major ideas behind our model is that the spatial and object pathways must be jointly active and learning to generate predictions about what will happen next. A related idea was proposed by Hinton (1981), who advocated solving the joint spatial configuration and object identification problems at the same time, with the goal of producing a canonical object representation that would then be easier to recognize.

However, the ill-posed and very high-dimensional nature of this problem proved intractable. Our approach avoids these problems by *first* developing the spatial prediction pathway independent of object recognition, using abstracted spatial blob representations, which is entirely tractable and easily learned. Then, we do not require a canonical object representation, but rather rely on well-established principles of hierarchical topographic connectivity to develop invariant object representations in the high levels of the *What* pathway (Fukushima, 1980; Riesenhuber & Poggio, 1999; O'Reilly et al., 2013).

Mumford's Models

David Mumford's early theoretical papers on the thalamus and cortex come the closest overall to capturing the central ideas in the current model, including the notion of the pulvinar as a kind of blackboard (Mumford, 1991) and the cortex as a generative model (Mumford, 1992). Although we only read these papers after developing our model, and there are many important differences in our approaches, the degree of concordance at the big-picture level is nevertheless remarkable.

Broader Implications of our Framework

Next, we consider a few of the most important broader implications of our framework.

Nature vs. Nurture in Development

There are many important developmental implications for a predictive learning approach in general (e.g., Elman, Bates, Karmiloff-Smith, Johnson, Parisi, & Plunkett, 1996; Munakata, McClelland, Johnson, & Siegler, 1997), and, as noted above, for the specific developmental requirements of our what-where-integration model. First, if you have a learning process that operates at a rate of 10 times per second, then a great deal of learning can accumulate very quickly. For example, the full sequence of training used in our model would represent just 21 hours of real-time learning at this rate. Of course, real-world environmental events may not be quite as dense a source of learning opportunities, and babies are certainly not awake very much at the start, but nevertheless it seems likely that a huge amount of predictive learning could be acquired by 4 months, when various studies indicate that babies have a decent understanding of basic physics (e.g., Spelke, 1994; Kellman & Spelke, 1983). Thus, this knowledge, which has been characterized as innate core knowledge (Spelke, 1994), may well be better described as learned. Nevertheless, given the ubiquitous nature of physics, coupled with genetically-coded learning mechanisms and developmental wiring processes, it is likely inevitable that all neurologically-intact babies will develop the same systematic predictive knowledge of this basic physics, so for all practical purposes, it certainly seems to be innate. Thus, the utility of simplistic nature vs. nurture dichotomies must be entirely rethought in the context of strong interactions be-

tween genetically-specified features of the brain and experience-expectant learning mechanisms (e.g., Elman et al., 1996; Greenough, Black, & Wallace, 1987).

In any case, there is now a great opportunity to explore more detailed data on the development of visual expectations about the world, using a more advanced version of our model and environment that contains multi-body interactions of various types (collisions, support, occlusion, etc). Furthermore, as noted above, the object representations learned by our model likely provide the foundation for subsequent word learning, and there is a large and somewhat contentious literature on this topic, which a more advanced multi-modal version of our model could hopefully contribute to (e.g., Stevens, Gleitman, Trueswell, & Yang, 2017; Yu & Smith, 2012; Colunga & Smith, 2005; Waxman & Gelman, 2009). This area is especially ripe for such models given a recent emphasis on collecting real-world experience samples that provide considerable insight and constraints (Yu & Smith, 2012; Roy, Frank, DeCamp, Miller, & Roy, 2015; Stevens et al., 2017).

Consciousness and Qualia

There are some potentially important implications of our framework for understanding the nature of consciousness, and what it feels like to be conscious of the visual world (qualia). The pulvinar plays a central role in our model as a kind of *projection screen*, but this naturally raises the question: is "anyone" watching this screen? Indeed, subjectively, there is a widespread seductive feeling that our brains have a kind of theater where the conscious part watches the incoming reports from the senses. Dennett (1991) refers to this as the *Cartesian Theater*, to deride the implicit dualism present in many theories (i.e., between the conscious part that watches the screen, and the unconscious part that projects representations onto it). But what if our brains really do have a kind of "silver screen of the Cartesian Theater" in the pulvinar (updating at film-appropriate alpha frame rates no less!), which underlies this pervasive subjective feeling of there being a kind of internal movie screen in our minds?

Without adopting any form of materialistic dualism, it is still possible that the pulvinar can play a critical role in organizing and coordinating diffuse brain areas around a common focus on the collaboratively-generated prediction of what will happen next. In so doing, we could say that this naturally contributes to the unitary nature of conscious experience, and provides a plausible substrate for how many different brain areas can share in a common perceptual-level sensory "qualia", which, because it is so strongly anchored by low-level visual areas (V1, V2), would have a distinctly "visual" feel to it. This kind of architecture would seem likely to produce a different emergent subjective experience than one where each area only interacts with its nearest neighbors, and is thus more "isolated" (higher-order areas in particular would be more strongly detached from low-level sensory details). This may also explain some of the mechanisms

behind an embodied, sensory-motor foundation to higher-level cognitive function (Barsalou, 2008, 2009; Anderson, 2003).

Critically, we avoid any strong localization of consciousness by virtue of the fact that each brain area is both a contributor to, and receiver of, this pulvinar projection screen, so there is no dualism of the form targeted by the Cartesian Theater notion — consciousness remains an emergent process characterized by coordination of processing across diffuse brain areas, which is a common notion across many different accounts (Baars, 1983, 2002; Dehaene & Naccache, 2001; Crick & Koch, 2003; Tononi, 2004; Lamme, 2006; Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008). In particular, the pulvinar may represent a different kind of global workspace than other accounts have postulated (Baars, 2002; Dehaene & Naccache, 2001), but with perhaps similar functional implications.

Finally, it is essential to recognize that consciousness is *inescapably dualist* — it is a property of *subjective* experience, which can *never* be described in purely *objective* terms. This is not substance dualism, but rather *perspective dualism* — it is literally definitionally impossible to transplant yourself into (another) human brain (you would become the other person, with no trace of yourself left, or some weird hybrid that is neither), so unless you happen to already be a human brain, you'll *never* know subjectively what it feels like to be one (and likewise for one individual brain to the next). This perspective dualism likely accounts for much of what is attributed to the *hard problem* of consciousness (Chalmers, 1995), without requiring any kind of substance dualism, and without preventing the attempt to map objective properties of the brain onto the subjective nature of experience. For example, it would be really interesting if we could selectively deactivate the pulvinar and subjectively report the effect on the nature of our experience. But that report would not enable others to actually experience the same thing, in the same way that attempting to convey the feeling of being on LSD or other powerful drugs is ultimately insufficient (no matter how poetic you get), if you haven't tried them yourself (and even then, you only truly know your own experience). Thus, while it is impossible to prove, the image of all these brain areas gathered around the silver screen of the pulvinar may underlie some important aspects of our subjective experience, and hence the seductive pull of the Cartesian Theater notion.

Predictions

A paper on the importance of predictive learning certainly must include a section on predictions from this framework! As in predictive learning, enumerating predictions from a theory provides a way of testing internal representations and refining them in light of observed data. There are so many possible predictions from our framework, and a good deal of the existing data has already been discussed above, so here we

highlight a few of the most central tests.

- Early developmental damage to the pulvinar should massively impair visual learning, but similar damage after developmental learning is complete should mainly affect attention (and also carefully-constructed learning tests that require learning in affected visual areas).

- Early developmental damage to MT (and probably DP) should paradoxically impair object recognition, by interfering with the partitioning of prediction error, but later in development the stabilized *What* pathway representations should be much less affected. The same applies to area LIP, but that might have even broader direct impairments that make it difficult to interpret. Given the relative homogeneity and plasticity of neocortex, other areas might be able to partially compensate, so this could be challenging to test effectively.

- The quantitative differences in response properties in the *What * Where* vs. *What* pathways as shown in Figure 11 and Table 1 should be testable using sufficiently large samples of neural recordings. The fuller integration in the *What * Where* pathway may emerge in the areas above MT (DP, MST, V6) in the larger scale context of the primate brain compared to our small-scale model.

- If it were possible to selectively block the 5IB intrinsic bursting neurons, or perhaps disable their bursting behavior in some other way, we would predict that this would have a significant impact on any task requiring temporal integration of information over time. For example, discriminating different individuals based on their walking motion, or recognizing a musical tune. More generally, if any person was brave enough to attempt taking a pharmacological agent that selectively interfered with 5IB bursting, we would predict that it would significantly disrupt the basic continuity of consciousness — everything would feel more fragmented and discontinuous and incoherent. Indeed, perhaps certain existing psychoactive substances can be understood in part in terms of their modulation of alpha bursting?

- Neocortical learning should also be significantly impaired with blockage of 5IB intrinsic bursting dynamics, because these contribute to the hypothesized plus phase of learning. To test this prediction, the widely-used statistical learning paradigm would be ideal, where sequences of tones or visual stimuli are presented, with various forms of statistical regularities (e.g., Aslin, Saffran, & Newport, 1998).

- Using large-scale lamina-specific neural recording techniques, it should be possible to quantify the information encoded in the layer 6 regular spiking (RS) neurons just after 5IB bursting, compared

to the information in the superficial layers just prior. Because we think that the layer 6 RS neurons convey the temporal context information from the prior alpha cycle, these two should be more strongly correlated in their information content, as compared to for example the information in superficial layers during the subsequent alpha cycle. Also, these layer 6 neurons should exhibit more rapid representational changes immediately post 5IB bursting compared to later in the cycle.

- A critical and only indirectly supported (Lim et al., 2015; Jedlicka et al., 2015; Zenke et al., 2017) property of our synaptic plasticity mechanism is the rapid updating of the plasticity threshold determining the boundary between LTD and LTP at the alpha time scale (as compared to the slower adaptation assumed in the BCM algorithm) — this could be tested much more directly using standard *in vitro* techniques. However, there may be important features of the awake *in vivo* environment that are essential for how the learning actually works, so that would be the ideal and only definitive test environment. Potentially modern optogenetic and imaging techniques would be capable of addressing this question.

- It should take at least two alpha cycles to process information from a new, exploratory fixation in a complex visual scene — the first alpha cycle will only have weak predictive and attentional deep layer representations associated with it, so a second one is required to generate a reliable prediction and more refined attentional spotlight. Thus, we predict that the modal fixation time in such cases should be around 200 msec. We are unsure of what may happen with more complex, novel, or otherwise hard-to-process stimuli: they may require more alpha cycles, or the duration of settling within a given alpha cycle may be stretched out as the constraint satisfaction process converges (Wyatte et al., 2012a).

- Instead of computing stable, static representations, the constant predictive pressure in this framework should favor rapidly-updating, dynamic representations that track the environment closely. For example, working memory representations of spatial locations may be encoded in retinotopic coordinates, and updated with every saccade, instead of using a more allocentric representation that does not require this updating (Wurtz, 2008; Cavanagh et al., 2010; Fix, Rougier, & Alexandre, 2011). This dynamic, constantly-updating, environmentally-tied vision of cognition is generally compatible with the embodied cognition approaches (Barsalou, 2008, 2009; Anderson, 2003; Smith & Thelen, 2003).

Unresolved Issues and Future Research

We have mentioned a number of unresolved issues and future directions throughout the paper. Here we highlight a few of the most important.

- Scaling up: How will the current model scale up to realistic 3D objects, larger spatial scales (allowing a difference between microsaccades and regular saccades), binocular and color vision, etc? We are confident in the basic principles, but much hard computational work remains to scale up the model to handle more realistic visual inputs, including likely adding additional high-level areas to specialize on encoding the relevant new dimensions in an efficient, systematic, and compact manner (e.g., CIP, next to LIP, appears to be specialized for 3D shape information, and interacts with the IT *What* pathway; Freud, Plaut, & Behrmann, 2016; Dromme, Premereur, Verhoef, Vanduffel, & Janssen, 2016; Tsutsui, Jiang, Yara, Sakata, & Taira, 2001).

- Scaling n: The attentional properties of our framework are only relevant in cluttered scenes with multiple different objects that could be tracked — these kinds of complex environments also need to be explored for many basic physical phenomena (collisions, support, occlusions etc). Will the LIP spatial blob representations provide a central organizing "FINST" pointer that coordinates attention and prediction across multiple brain areas, for the attentionally selected objects (Pylyshyn, 1989; Cavanagh et al., 2010; O'Reilly et al., 2014b)?

- Scaling out: how does visual predictive learning interact with auditory and/or somatosensory predictive learning? As noted earlier, including auditory inputs is essential for exploring language learning, and forward-model-like predictive learning in speech, and motor control more broadly.

- Scaling on: how do predictions and representations of longer time-scale events and episodes build upon the fast alpha-rhythm sensory predictive learning loop? We noted that the medial temporal lobe can encode two alpha trials in one of its characteristic theta cycles, but how are yet longer time scales encoded? Robust active maintenance in the prefrontal cortex likely plays a critical role, but how are its representations trained in the context of predictive learning?

- Biological and mechanical motion: living things and machines move differently than inert physical objects — if we are to accurately predict the visual world, a strong interaction between *What* * *Where* is necessary for these things. From the principles of our framework, we would predict that that a specialized higher-level area above the basic *What* * *Where* pathway, with strong input from the *What* pathway, would be needed to learn these higher-order cases, and indeed just such an area in the STS, anatomically above the MT, MST pathway, has been identified (Puce & Perrett, 2003).

- Dynamic alpha: there is considerable evidence that the alpha rhythm can be entrained by external stimuli, which is important for ensuring that the temporal context updates track relevant events in the

environment. The current model just uses a fixed trial timing, so relevant mechanisms to support alpha phase entrainment need to be incorporated into our model.

- Dynamic activations: As reviewed above, there are many short-time-scale dynamics that may play an important role in shaping the time-evolution of neural representations at the alpha time scale — these may affect the dynamics of prediction updating in important ways and should be thoroughly explored.

- To what extent do the lessons from our pulvinar-based model apply to the LGN, in its interconnectivity with the retina and V1? A fundamental difference is that there are no alpha-bursting plus phase driver inputs to the LGN as far as we know, so it would seem that the V1 / LGN system learns in a purely Hebbian manner without the benefit of predictive error signals, which is consistent with many Hebbian models of V1 learning (e.g., Miller, Keller, & Stryker, 1989; Bednar & Miikkulainen, 2003). However, the same prediction-generation pathway from layer 6CT to LGN does exist — likely this is playing a largely attentional role as it also plays in the attentional aspect of our model. Nevertheless, all of these issues bear deeper reexamination to see if there might be some other interesting kinds of thalamocortical learning dynamics taking place, which would likely also apply to other modalities.

Conclusions

In conclusion, our model clearly builds on ideas that have long been advocated in understanding neocortical function, while also adding some important new elements, that together have produced a coherent, functional, first pass working model demonstrating the sufficiency of the framework to achieve significant forms of learning through the predictive mechanism. There are many outstanding questions still, so a pessimist may not yet be convinced of the value of this framework, and certainly we have a tremendous amount left to learn. Finally, it is worth observing that the odds of discovering a model of this complexity through a purely bottom up, empirically-driven approach seem rather small. Similarly, purely computational or cognitive-level theorists would probably not have arrived at some of the key insights provided by the biology. Thus, a systems-focused, computational-modeling approach that integrates elements from all these different levels of analysis can play a critical role in advancing our understanding of the complexities of brain function.

References

Abbott, L. F., Varela, J. A., Sen, K., & Nelson, S. B. (1997). Synaptic Depression and Cortical Gain Control. *Science*, *275*, 220.

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

Aisa, B., Mingus, B., & O'Reilly, R. (2008). The emergent neural modeling system. *Neural Networks*, *21*(8), 1146–1152.

Anderson, B., & Sheinberg, D. L. (2008). Effects of temporal context and temporal expectancy on neural activity in inferior temporal cortex. *Neuropsychologia*, *46*(4), 947–957.

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought.* Mahwah, NJ: Erlbaum.

Anderson, M. L. (2003). Embodied Cognition: A field guide. *Artificial Intelligence*.

Artola, A., Bröcher, B., & Singer, W. (1990). Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. *Nature*, *347*(6288), 69–72.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of Conditional Probability Statistics By 8-Month-Old Infants. *Psychological Science*, *9*(4), 321–324.

Baars, B. J. (1983). Conscious Contents Provide the Nervous System with Coherent, Global Information. In R. J. Davidson, G. E. Schwartz, & D. H. Shapiro (Eds.), *Consciousness and Self-Regulation* (pp. 41–79). Plenum.

Baars, B. J. (2002). The conscious access hypothesis: Origins and recent evidence. *Trends in cognitive sciences*, *6*, 47–52.

Barlow, H. B. (1989). Unsupervised Learning. *Neural Computation*, *1*, 295–311.

Barsalou, L. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B*, *364*(1521), 1281–1289.

Barsalou, L. W. (2008). Grounded cognition. *Annual review of psychology*, *59*, 617–645.

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.

Bear, M. F., & Malenka, R. C. (1994). Synaptic Plasticity: LTP and LTD. *Current Opinion in Neurobiology*, *4*, 389–399.

Bednar, J. A., & Miikkulainen, R. (2003). Self-organization of spatiotemporal receptive fields and laterally connected direction and orientation maps. *Neurocomputing*, *52*, 473–480.

Behrmann, M., & Plaut, D. C. (2013). Distributed circuits, not circumscribed centers, mediate visual recognition. *Trends in Cognitive Sciences*, *17*(5), 210–219.

Bender, D. B. (1981). Retinotopic organization of macaque pulvinar. *Journal of Neurophysiology*, *46*(3), 672–693.

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

Bengio, Y., Courville, A., & Vincent, P. (2013a). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

Bengio, Y., Yao, L., Alain, G., & Vincent, P. (2013b). Generalized Denoising Auto-Encoders as Generative Models. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 899–907). Curran Associates, Inc.

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.

Bollimunta, A., Chen, Y., Schroeder, C. E., & Ding, M. (2008). Neuronal mechanisms of cortical alpha oscillations in awake-behaving macaques. *The Journal of Neuroscience*, *28*(40), 9976–9988.

Bollimunta, A., Mo, J., Schroeder, C. E., & Ding, M. (2011). Neuronal mechanisms and attentional modulation of corticothalamic alpha oscillations. *The Journal of Neuroscience*, *31*(13), 4935–4943.

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

Bourne, J. A., & Rosa, M. G. P. (2006). 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.

Brette, R., & Gerstner, W. (2005). Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of Neurophysiology*, *94*(5), 3637–3642.

Bridge, H., Leopold, D. A., & Bourne, J. A. (2016). Adaptive Pulvinar Circuitry Supports Visual Cognition. *Trends in Cognitive Sciences*, *20*(2), 146–157.

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.

Burgess, N., & O'Keefe, J. (1997). Neuronal computations underlying the firing of place cells and their role in navigation. *Hippocampus*, *6*, 749–762.

Buschman, T. J., Siegel, M., Roy, J. E., & Miller, E. K. (2011). Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of Sciences*, *108*(27), 11252–11255.

Buxhoeveden, D. P., & Casanova, M. F. (2002). The minicolumn hypothesis in neuroscience. *Brain*, *125*(Pt 5), 935–951.

Carpenter, G., & Grossberg, S. (1987). A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine. *Computer Vision, Graphics, and Image Processing*, *37*(1), 54–115.

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.

Cer, D., & O'Reilly, R. C. (2006). Neural mechanisms of binding in the hippocampus and neocortex: Insights from computational models. In H. D. Zimmer, A. Mecklinger, & U. Lindenberger (Eds.), *Handbook of binding & memory. Perspectives from cognitive neuroscience*. Oxford: Oxford University Press.

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *3(1)*, 200–217.

Ciresan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column Deep Neural Networks for Image Classification. *IEEE Conf. on Computer Vision and Pattern Recognition CVPR 2012*, 3642–3649.

Ciresan, D. C., Meier, U., Gambardella, L. M., & Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. *Neural Computation*, *22*(12), 3207–3220.

Cleeremans, A. (1993). *Mechanisms of Implicit Learning: Connectionist models of sequence processing*. Cambridge, MA: MIT Press.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite State Automata and Simple Recurrent Networks. *Neural Computation*, *1*(3), 372–381.

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.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological review*, *112*(2), 347–382.

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

Cooper, L. N., Intrator, N., Blais, B. S., & Shouval, H. (2004). *Theory of Cortical Plasticity.* New Jersey: World Scientific.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

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.

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*, 129–132.

Crick, F., & Koch, C. (2003). A framework for consciousness. *Nature Neuroscience*, *6*(2), 119–126.

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

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*(1-2), 1–37.

Dennett, D. C. (1991). *Consciousness Explained.* Boston and London: Little, Brown, and Co.

Desimone, R. (1996). Neural Mechanisms for Visual Memory and Their Role in Attention. *Proceedings of the National Academy of Sciences*, *93*(24), 13494–13499.

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*, 193–222.

Dougherty, K., Cox, M. A., Ninomiya, T., Leopold, D. A., & Maier, A. (2017). Ongoing Alpha Activity in V1 Regulates Visually Driven Spiking Responses. *Cerebral Cortex*, *27*(2), 1113–1124.

Douglas, R. J., & Martin, K. A. C. (2004). Neuronal Circuits of the Neocortex. *Annual Review of Neuroscience*, *27*, 419–451.

Dromme, I. C. V., Premereur, E., Verhoef, B.-E., Vanduffel, W., & Janssen, P. (2016). Posterior Parietal Cortex Drives Inferotemporal Activations During Three-Dimensional Object Vision. *PLOS Biology*, *14*(4), e1002445.

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.

Elman, J. L. (1991). Distributed Representations, Simple Recurrent Networks, and Grammatical Structure. *Machine Learning*, *7*(2-3), 195–225.

Engel, A. K., König, P., Kreiter, A. K., Schillen, T. B., & Singer, W. (1992). Temporal coding in the visual cortex: New vistas on integration in the nervous system. *Trends in neurosciences*, *15*(6), 218–226.

Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. F. (2008). The spatiotemporal profile of cortical processing leading up to visual perception. *Journal of Vision*, *8*(1), 1–12.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex*, *1*(1), 1–47.

Field, D. J. (1994). What Is the Goal of Sensory Coding? *Neural Computation*, *6*(4), 559–601.

Fix, J., Rougier, N., & Alexandre, F. (2011). A dynamic neural field approach to the covert and overt deployment of spatial attention. *Cognitive Computation*, *3*(1), 279–293.

Flint, A. C., & Connors, B. W. (1996). Two types of network oscillations in neocortex mediated by distinct glutamate receptor subtypes and neuronal populations. *Journal of Neurophysiology*, *75*(2), 951–957.

Foldiak, P. (1991). Learning Invariance from Transformation Sequences. *Neural Computation*, *3*(2), 194–200.

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.

Freud, E., Plaut, D. C., & Behrmann, M. (2016). 'What' is happening in the dorsal visual pathway. *Trends in Cognitive Sciences*, *20*(10), 773–784.

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

Friston, K. (2008). Hierarchical Models in the Brain. *PLOS Computational Biology*, *4*(11), e1000211.

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

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74.

George, D., & Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits. *PLoS Computational Biology*, *5*(10).

Gerstner, W., & Naud, R. (2009). How Good Are Neuron Models? *Science*, *326*(5951), 379–380.

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

Gray, C. M., Engel, A. K., König, P., & Singer, W. (1992). Synchronization of oscillatory neuronal responses in cat striate cortex: Temporal properties. *Visual neuroscience*, *8*, 337–347.

Greenough, W. T., Black, J. E., & Wallace, C. S. (1987). Experience and Brain Development. *Child Development*, *58*, 539–559.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23.

Guenther, F. H., & Vladusich, T. (2012). A Neural Theory of Speech Acquisition and Production. *Journal of neurolinguistics*, *25*.

Haegens, S., Ncher, V., Luna, R., Romo, R., & Jensen, O. (2011). -Oscillations in the monkey sensorimotor network influence discrimination performance by rhythmical inhibition of neuronal spiking. *Proceedings of the National Academy of Sciences USA*, *108*(48), 19377–19382.

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

Hennig, M. H. (2013). Theoretical models of synaptic short term plasticity. *Frontiers in Computational Neuroscience*, *7*.

Herbort, O., & Butz, M. V. (2012). Too Good to be True? Ideomotor Theory from a Computational Perspective. *Frontiers in psychology*, *3*.

Hinton, G. E. (1981, January). A Parallel Computation That Assigns Canonical Object-Based Frames of Reference. *Proceedings of the 7th IJCAI* (pp. 683–685). Vancouver.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, *14*, 1771–1800.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in cognitive sciences*, *11*(10), 428–434.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed Representations. In D. E. Rumelhart, J. L. McClelland, & P. R. Group (Eds.), *Parallel Distributed Processing. Volume 1: Foundations* (pp. 77–109). Cambridge, MA: MIT Press.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in cognitive sciences*, *8*(11), 494–500.

Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*(4), 613–622.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(8), 2554–2558.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences of the United States of America*, *81*, 3088–3092.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological review*, *99*(3), 480–517.

Idiart, M. A. P., & Lisman, J. E. (1995). Storage of 7 pm 2 Short-Term Memories in Oscillatory Subcycles. *Science*, *267*, 1512.

James, W. (1890). *The Principles of Psychology*. New York: Henry Holt.

Jedlicka, P., Benuskova, L., & Abraham, W. C. (2015). A Voltage-Based STDP Rule Combined with Fast BCM-Like Metaplasticity Accounts for LTP and Concurrent "Heterosynaptic" LTD in the Dentate Gyrus In Vivo. *PLOS Computational Biology*, *11*(11), e1004588.

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

Jilk, D., Lebiere, C., O'Reilly, R., & Anderson, J. (2008). SAL: An explicitly pluralistic cognitive architecture. *Journal of Experimental & Theoretical Artificial Intelligence*, *20*(3), 197–218.

Johnson, R. R., & Burkhalter, A. (1997). A polysynaptic feedback circuit in rat visual cortex. *The Journal of Neuroscience*, *17*(18), 7129–7140.

Jordan, M. I. (1989). Serial Order: A Parallel, Distributed Processing Approach. In J. L. Elman, & D. E. Rumelhart (Eds.), *Advances in Connectionist Theory: Speech*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward Models: Supervised Learning with a Distal Teacher. *Cognitive Science*, *16*(3), 307–354.

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

Kanerva, P. (1988). *Sparse Distributed Memory*. Boston: Bradford MIT.

Kawato, M., Furukawa, K., & Suzuki, R. (1987). A hierarchical neural-network model for control and learning of voluntary movement. *Biological cybernetics*, *57*.

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.

Kellman, P. J., & Spelke, E. (1983). Perception of partially occluded objects in infancy. *Cognitive Psychology*, *15*(4), 483–524.

Ketz, N., Morkonda, S. G., & O'Reilly, R. C. (2013). Theta coordinated error-driven learning in the hippocampus. *PLoS Computational Biology*, *9*, e1003067.

Ketz, N. A., Jensen, O., & O'Reilly, R. C. (2015). Thalamic pathways underlying prefrontal cortex-medial temporal lobe oscillatory interactions. *Trends in neurosciences*, *38*, 3–12.

Kiorpes, L., Price, T., Hall-Haro, C., & Anthony Movshon, J. (2012). Development of sensitivity to global form and motion in macaque monkeys (Macaca nemestrina). *Vision Research*, *63*, 34–42.

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.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 1097–1105). Curran Associates, Inc.

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.

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of Neuronal Oscillations as a Mechanism of Attentional Selection. *Science*, *320*(5872), 110–113.

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494–501.

Le, Q. V., Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J., & Ng, A. Y. (2012). Building high-level features using large scale unsupervised learning. *In ICML*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. *Advances in Neural Information Processing Systems* (pp. 396–404). Morgan Kaufmann.

Lee, T. S. (2015). The Visual System's Internal Model of the World. *Proceedings of the IEEE*, *103*(8), 1359–1378.

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

Lee, T. S., & Nguyen, M. (2001). Dynamics of subjective contour formation in the early visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 1907–1911.

Lee, T. S., Yang, C. F., Romero, R. D., & Mumford, D. (2002). Neural activity in early visual cortex reflects behavioral experience and higher-order perceptual saliency. *Nature Neuroscience*, *5*(6), 589–597.

Leonard, K., & Yorton, T. (2015). *Yes, and: How Improvisation Reverses 'no, But' Thinking and Improves Creativity and Collaboration–lessons from the Second City*. Harper Collins.

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.

Lisman, J. (1990). A mechanism for the Hebb and the anti-Hebb processes underlying learning and memory. *Proceedings of the National Academy of Sciences USA*, *86*(23), 9574–9578.

Lisman, J. (1995). The CaM kinase II hypothesis for the storage of synaptic memory. *Trends in neurosciences*, *17*, 406.

Lisman, J., & Redish, A. D. (2009). Prediction, sequences and the hippocampus. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1193–1201.

Lisman, J. E., & Jensen, O. (2013). The theta-gamma neural code. *Neuron*, *77*(6), 1002–16.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual Object Recognition. *Annual Review of Neuroscience*, *19*, 577–621.

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.

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.

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *The Journal of Neuroscience*, *35*(39), 13402–13418.

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

Markov, N. T., Ercsey-Ravasz, M. M., Gomes, R., R, A., Lamy, C., Magrou, L., Vezoli, J., Misery, P., Falchier, A., Quilodran, R., Gariel, M. A., Sallet, J., Gamanut, R., Huissoud, C., Clavagnier, S., Giroud, P., Sappey-Marinier, D., Barone, P., Dehay, C., Toroczkai, Z., Knoblauch, K., Van Essen, D. C., & Kennedy, H. (2014a). A Weighted and Directed Interareal Connectivity Matrix for Macaque Cerebral Cortex. *Cerebral Cortex*, *24*(1), 17–36.

Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., Lamy, C., Misery, P., Giroud, P., Ullman, S., Barone, P., Dehay, C., Knoblauch, K., & Kennedy, H. (2014b). Anatomy of hier-

archy: Feedforward and feedback pathways in macaque visual cortex: Cortical counterstreams. *Journal of Comparative Neurology*, *522*(1), 225–259.

Markram, H., & Tsodyks, M. (1996). Redistribution of synaptic efficacy between neocortical pyramidal neurons. *Nature*, *382*(6594), 807–810.

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.

Miall, R. C., & Wolpert, D. M. (1996). Forward Models for Physiological Motor Control. *Neural Netw*, *9*(8), 1265–1279.

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.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates, Inc.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Miller, K. D., Keller, J. B., & Stryker, M. P. (1989). Ocular dominance column development: Analysis and simulation. *Science (New York, N.Y.)*, *245*, 605–615.

Morton, J., & Johnson, M. H. (1991). CONSPEC and CONLERN: A two-process theory of infant face recognition. *Psychological review*, *98*, 164–181.

Mountcastle, V. B. (1957). Modality and topographic properties of single neurons of cat's somatic sensory cortex. *Journal of Neurophysiology*, *20*(4), 408–434.

Mountcastle, V. B. (1997). The columnar organization of the neocortex. *Brain*, *120*(Pt 4), 701–722.

Movellan, J. R. (1990, January). Contrastive Hebbian Learning in the Continuous Hopfield Model. In D. S. Touretzky, G. E. Hinton, & T. J. Sejnowski (Eds.), *Proceedings of the 1989 Connectionist Models Summer School* (pp. 10–17). San Mateo, CA: Morgan Kaufman.

Müller, J. R., Metha, A. B., Krauskopf, J., & Lennie, P. (1999). Rapid adaptation in visual cortex to the structure of images. *Science (New York, N.Y.)*, *285*, 1405.

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.

Munakata, Y., McClelland, J. L., Johnson, M. H., & Siegler, R. S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review*, *104*, 686–713.

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.

Neupane, S., Guitton, D., & Pack, C. C. (2017). Coherent alpha oscillations link current and future receptive fields during saccades. *Proceedings of the National Academy of Sciences*, 201701672.

Nishimura, M., Scherf, S., & Behrmann, M. (2009). Development of object recognition in humans. *F1000 Biology Reports*, *1*.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, *110*(4), 611–646.

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. *The Journal of neuroscience*, *33*(5).

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

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, *37*(23), 3311–3325.

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. (2001). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation*, *13*(6), 1199–1242.

O'Reilly, R. C., & Busby, R. S. (2002, January). Generalizable Relational Binding from Coarse-coded Distributed Representations. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems (NIPS) 14*. Cambridge, MA: MIT Press.

O'Reilly, R. C., Busby, R. S., & Soto, R. (2003). Three Forms of Binding and their Neural Substrates: Alternatives to Temporal Synchrony. In A. Cleeremans (Ed.), *The Unity of Consciousness: Binding, Integration, and Dissociation* (pp. 168–192). Oxford: Oxford University Press.

O'Reilly, R. C., Hazy, T. E., & Herd, S. A. (2016). The Leabra cognitive architecture: How to play 20 principles with nature and win! In S. Chipman (Ed.), *Oxford handbook of cognitive science*. Oxford University Press.

O'Reilly, R. C., Hazy, T. E., Mollick, J., Mackie, P., & Herd, S. (2014a). Goal-Driven Cognition in the Brain: A Computational Framework. *arXiv:1404.7591 [q-bio]*.

O'Reilly, R. C., & Johnson, M. H. (1994). Object Recognition and Sensitive Periods: A Computational Analysis of Visual Imprinting. *Neural Computation*, *6*(3), 357–389.

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., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors (2012). *Computational Cognitive Neuroscience*. Wiki Book, 1st Edition, URL: http://ccnbook.colorado.edu.

O'Reilly, R. C., Petrov, A. A., Cohen, J. D., Lebiere, C. J., Herd, S. A., & Kriete, T. (2014b). 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).

Pastalkova, E., Itskov, V., Amarasingham, A., & Buzsáki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. *Science (New York, N.Y.)*, *321*(5894), 1322–1327.

Pezzulo, G., & Castelfranchi, C. (2009). Thinking as the control of imagination: A conceptual framework for goal-directed systems. *Psychological research*, *73*.

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

Pouget, A., & Sejnowski, T. J. (1997). A new view of hemineglect based on the response properties of parietal neurones. *Philosophical Transactions of the Royal Society of London B Biol Sci*, *352*(1360), 1449–1459.

Puce, A., & Perrett, D. (2003). Electrophysiology and brain imaging of biological motion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*(1431), 435–445.

Pylyshyn, Z. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, *32*(1), 65–97.

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.

Rao, S. G., Williams, G. V., & Goldman-Rakic, P. S. (1999). Isodirectional tuning of adjacent interneurons and pyramidal cells during working memory: Evidence for microcolumnar organization in PFC. *Journal of Neurophysiology*, *81*(4), 1903–1916.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., & Raiko, T. (2015). Semi-supervised Learning with Ladder Networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28* (pp. 3546–3554). Curran Associates, Inc.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

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.

Rockland, K. S., & Pandya, D. N. (1979). Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Research*, *179*(1), 3–20.

Rodman, H. R. (1994). Development of Inferior Temporal Cortex in the Monkey. *Cerebral Cortex*, *4*(5), 484–498.

Rodriguez, A., Whitson, J., & Granger, R. (2004). Derivation and analysis of basic computational operations of thalamocortical circuits. *Journal of Cognitive Neuroscience*, *16*(5), 856–877.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663–12668.

Rumelhart, D. E., & McClelland, J. L. (1982). 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.

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.

Scellier, B., & Bengio, Y. (2017). Equilibrium Propagation: Bridging the Gap between Energy-Based Models and Backpropagation. *Frontiers in Computational Neuroscience*, *11*.

Schubert, D., Kotter, R., & Staiger, J. F. (2007). Mapping functional connectivity in barrel-related columns reveals layer- and cell type-specific microcircuits. *Brain Structure & Function*, *212*(2), 107–119.

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: Relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences*, *12*(8), 314–321.

Shadmehr, R. (2017). Learning to Predict and Control the Physics of Our Movements. *Journal of Neuroscience*, *37*(7), 1663–1671.

Shao, Z., & Burkhalter, A. (1996). Different Balance of Excitation and Inhibition in Forward and Feedback Circuits of Rat Visual Cortex. *The Journal of Neuroscience*, *16*(22), 7353–7365.

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

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

Shouval, H. Z., Wang, S. S.-H., & Wittenberg, G. M. (2010). Spike timing dependent plasticity: A consequence of more fundamental learning rules. *Frontiers in Computational Neuroscience*, *4*(19).

Silva, L. R., Amitai, Y., & Connors, B. W. (1991). Intrinsic oscillations of neocortex generated by layer 5 pyramidal neurons. *Science*, *251*(4992), 432–435.

Sincich, L. C., Park, K. F., Wohlgemuth, M. J., & Horton, J. C. (2004). Bypassing V1: a direct geniculate input to area MT. *Nature Neuroscience*, *7*(10), 1123–1128.

Smith, L. B., & Thelen, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences*, *7*, 343–348.

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. (1994). Initial knowledge: Six suggestions. *Cognition*, *50*, 431–445.

Spratling, M. W. (2008). Reconciling predictive coding and biased competition models of cortical function. *Frontiers in Computational Neuroscience*, *2*(4), 1–8 (online).

Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017). The Pursuit of Word Meanings. *Cognitive Science*, *41*, 638–676.

Stocco, A., Lebiere, C., & Anderson, J. (2010). Conditional Routing of Information to the Cortex: A Model of the Basal Ganglia's Role in Cognitive Coordination. *Psychological Review*, *117*, 541–574.

Strick, P. L., Dum, R. P., & Fiez, J. A. (2009). Cerebellum and Nonmotor Function. *Annual Review of Neuroscience*, *32*(1), 413–434.

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.

Supèr, H., Spekreijse, H., & Lamme, V. A. (2001). Two distinct modes of sensory processing observed in monkey primary visual cortex (V1). *Nature Neuroscience*, *4*(3), 304–310.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, *19*, 109–139.

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.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*, 42.

Tsutsui, K.-I., Jiang, M., Yara, K., Sakata, H., & Taira, M. (2001). Integration of Perspective and Disparity Cues in Surface-Orientation–Selective Neurons of Area CIP. *Journal of Neurophysiology*, *86*(6), 2856–2867.

Ullman, S. (1995). Sequence seeking and counter streams: A computational model for bidirectional information flow in the visual cortex. *Cerebral cortex*, *5*(1), 1–11.

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.

Valpola, H. (2014). From neural PCA to deep unsupervised learning. *arXiv:1411.7783 [cs, stat]*.

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.

Verduzco-Flores, S. O., & O'Reilly, R. C. (2015). How the credit assignment problems in motor control could be solved after the cerebellum predicts increases in error. *Frontiers in Computational Neuroscience*, *9*.

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

von Holst, E. (1954). Relations between the central Nervous System and the peripheral organs. *The British Journal of Animal Behaviour*, *2*(3), 89–94.

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.

Wang, D., Buhmann, J., & von der Malsburg, C. (1990). Pattern Segmentation in Associative Memory. *Neural Computation*, *2*(1), 94–106.

Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, *13*(6), 258–263.

Wiggs, C. L., & Martin, A. (1998). Properties and mechanisms of perceptual priming. *Current Opinion in Neurobiology*, *8*(2), 227–233.

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.

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, *14*, 715–770.

Wolpert, D. M., Miall, R. C., & Kawato, M. (1998). Internal models in the cerebellum. *Trends in Cognitive Sciences*, *2*(9), 338–347.

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

Wyatte, D., Curran, T., & O'Reilly, R. (2012a). The limits of feedforward vision: Recurrent processing promotes robust object recognition when objects are degraded. *Journal of Cognitive Neuroscience*, *24*(11), 2248–2261.

Wyatte, D., Herd, S., Mingus, B., & O'Reilly, R. (2012b). The Role of Competitive Inhibition and Top-Down Feedback in Binding during Object Recognition. *Frontiers in Psychology*, *3*(182).

Xie, X., & Seung, H. S. (2003). Equivalence of backpropagation and Contrastive Hebbian Learning in a layered network. *Neural Computation*, *15*(2), 441–454.

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.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

Zemel, R. S., Williams, C. K. I., & Mozer, M. C. (1995). Lending Direction to Neural Networks. *Neural Networks*, *8*, 503.

Zenke, F., Gerstner, W., & Ganguli, S. (2017). The temporal paradox of Hebbian learning and homeostatic plasticity. *Current Opinion in Neurobiology*, *43*, 166–176.

Zipser, D., & Andersen, R. A. (1988). A Backpropagation Programmed Network That Simulates Response Properties of a Subset of Posterior Parietal Neurons. *Nature*, *331*, 679–684.

Zoccolan, D., Kouh, M., Poggio, T., & DiCarlo, J. J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *The Journal of neuroscience*, *27*(45).

Appendix: Computational Model Details

This appendix provides more information about the *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:

`http://grey.colorado.edu/CompCogNeuro/index.php/CCN_Repository`.

And 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://ccnbook.colorado.edu`.

Layer Sizes and Structure

Figure 3 shows the general configuration of the model, and Table 2 shows the specific sizes of each of the layers, and where they receive inputs from. The main text contains figures showing the patterns of connectivity, which establish the three pathways (Figures 7, 8).

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

Projections

Detailing each of the specific parameters associated with the different projections shown in Table 2 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 8x8 unit groups in V2 are reduced down to the 4x4 groups in V3 via a 4x4 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 current deep convolutional neural networks. However, we do not share weights across units as in a true convolutional network.

The projections from ObjVel (object velocity) and SaccadePlan layers to LIPs,d 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

		Units		Groups		
Area	**Name**	**X**	**Y**	**X**	**Y**	**Receiving Projections**
V1	V1s	4	4	8	8	
	V1p	4	4	8	8	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
	V2d	10	10	8	8	V2s V1p LIPd LIPp V3d V4d V3s TEOs
	V2p	10	10	8	8	V2s V3d V4d TEOd
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	V1s LIPd
V3	V3s	10	10	4	4	V2s V4s TEOs MTs LIPs V1p MTp TEOd
	V3d	10	10	4	4	V3s V1p MTp LIPd MTd V4d V4s MTs TEOs
	V3p	10	10	4	4	V3s V2d MTd TEOd
MT	MTs	10	10			V2s V3s TEOs V1p V3p TEOp OFCp
	MTd	10	10			MTs V1p MTp OFCp TEOd
	MTp	10	10			MTs V2d V3d MTd TEOd
V4	V4s	10	10	4	4	V2s TEOs V1p OFCp
	V4d	10	10	4	4	V4s V1p V4p OFCp TEOd TEOs
	V4p	10	10	4	4	V4s V2d V3d V4d TEOd
TEO	TEOs	8	8	4	4	V4s V1p
	TEOd	8	8	4	4	TEOs TEOd V1p V4p TEOp OFCp
	TEOp	8	8	4	4	TEOs V3d V4d TEOd

Table 2: Layer sizes, showing numbers of units in one unit group (or entire layer if Group is missing), and the number of Groups of such units, along X,Y axes. Each area has three associated layers: *s* = superficial layer, *d* = deep layer, *p* = pulvinar layer (driven by 5IB neurons from associated area).

faster convergence of the LIP *Where* pretraining compared to purely random initial weights, consistent with the basis function theory and related empirical observations (Zipser & Andersen, 1988; Pouget & Sejnowski, 1997).

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

As noted in the main text, 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 the dynamic visual environment as described in the main text. Because the start of each sequence of 4 trials 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.

The plots of learning trajectories have been smoothed with a gaussian kernel with a half-width of 8 epochs, sigma = 4 epochs, to make the different lines more easily discriminable — there is a reasonably high level of random noise in performance due to random variation in the environment parameters etc, so this smooths that out and allows the mean level to be visible.

Model Algorithms

The model was implemented using the Leabra framework, which is described in detail in previous publications (O'Reilly et al., 2016; O'Reilly et al., 2012; O'Reilly & Munakata, 2000; O'Reilly, 2001, 1998, 1996), and summarized here. The main implementation of Leabra is in the *emergent* software (Aisa, Mingus, & O'Reilly, 2008), and another detailed explanation of the algorithm, and simple implementations of all the equations in Python and MATLAB, are available from:

`https://grey.colorado.edu/emergent/index.php/Leabra`

These same equations and standard parameters have been used to simulate over 40 different models in O'Reilly et al. (2012) and O'Reilly and Munakata (2000), 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.

Leabra Algorithm Equations

The pseudocode for Leabra is given here, showing exactly how the pieces of the algorithm fit together, using the equations and variables from the actual code. The implementation contains a number of optimizations (including vectorization and GPU code), but this provides the core math in simple form.

See the `Matlab` directory in the emergent `svn` source directory for a complete implementation of these equations in Matlab, coded by Sergio Verduzco-Flores — this can be a lot simpler to read than the highly optimized C++ source code.

Timing

Leabra is organized around the following timing, based on an internally-generated alpha-frequency (10 Hz, 100 msec periods) cycle of expectation followed by outcome, supported by neocortical circuitry in the deep layers and the thalamus, as hypothesized in the `DeepLeabra` extension to standard Leabra:

- A **Trial** lasts 100 msec (10 Hz, alpha frequency), and comprises one sequence of expectation — outcome learning, organized into 4 quarters.
  - Biologically, the deep neocortical layers (layers 5, 6) and the thalamus have a natural oscillatory rhythm at the alpha frequency. Specific dynamics in these layers organize the cycle of expectation vs. outcome within the alpha cycle.
- A **Quarter** lasts 25 msec (40 Hz, gamma frequency) — the first 3 quarters (75 msec) form the expectation / minus phase, and the final quarter are the outcome / plus phase.
  - Biologically, the superficial neocortical layers (layers 2, 3) have a gamma frequency oscillation, supporting the quarter-level organization.
- A **Cycle** represents 1 msec of processing, where each neuron updates its membrane potential etc according to the above equations.

Variables

LeabraUnits are organized into LeabraLayers, which sometimes have unit groups (which are now typically purely virtual, not actual Unit_Group objects). The LeabraUnit has the following key parameters, along with a number of others that are used for other non-default algorithms and various optimizations, etc.

- **act** = activation sent to other units
- **act_nd** = non-depressed activation — prior to application of any short-term plasticity
- **net_raw** = raw netinput, prior to time-averaging
- **net** = time-averaged excitatory conductance (net input)
- **gc_i** = inhibitory conductance, computed from FFFB inhibition function typically
- **I_net** = net current, combining excitatory, inhibitory, and leak channels

- **v_m** = membrane potential
- **v_m_eq** = equilibrium membrane potential — not reset by spikes — just keeps integrating
- **adapt** = adaptation current
- **avg_ss** = super-short term running average activation
- **avg_s** = short-term running average activation, integrates over avg_ss, represents plus phase learning signal
- **avg_m** = medium-term running average activation, integrates over avg_s, represents minus phase learning signal
- **avg_l** = long-term running average activation, integrates over avg_m, drives long-term floating average for Hebbian learning
- **avg_l_lrn** = how much to use the avg_l-based Hebbian learning for this receiving unit's learning — in addition to the basic error-driven learning — this can optionally be dynamically updated based on the avg_l factor and average level of error in the receiving layer, so that this Hebbian learning constraint can be stronger as a unit gets too active and needs to be regulated more strongly, and in proportion to average error levels in the layer.
- **avg_s_eff** = effective avg_s value used in learning — includes a small fraction (.1) of the avg_m value, for reasons explained below.

Units are connected via synapses parameterized with the following variables. These are actually stored in an optimized vector format, but the LeabraCon object contains the variables as a template.

- **wt** = net effective synaptic weight between objects — subject to contrast enhancement compared to fwt and swt
- **dwt** = delta-wt — change in synaptic weights due to learning
- **dwavg** = time-averaged absolute value of weight change, for normalizing weight changes
- **moment** = momentum integration of weight changes
- **fwt** = fast weight — used for advanced fast and slow weight learning dynamic — otherwise equal to swt — stored as non-contrast enhanced value
- **swt** = slow weight — standard learning rate weight — stored as non-contrast enhanced value — optional

Activation Update Cycle (every 1 msec): Net input, Inhibition, Activation

For every cycle of activation updating, compute the net input, inhibition, membrane potential, and activation:

- **Net input** (see LeabraUnitSpec.cpp for code):
  - **net_raw** += (sum over recv connections of:) scale_eff * act * wt
    - **scale_eff** = https://grey.colorado.edu/emergent/index.php/Leabra_Netin_Scalin factor that includes 1/N to compute an average, plus wt_scale.rel and abs relative and absolute scaling terms.
    - **act** = sending unit activation
    - **wt** = receiving connection weight value between sender and receiver
    - does this very efficiently by using a sender-based computation, that only sends *changes* (deltas) in activation values — typically only a few percent of neurons send on any given cycle.
  - **net** += dt.integ * dt.net_dt * (net_raw - net)
    - time integration of net input, using net_dt (1/1.4 default), and global integration time constant, dt.integ (1 = 1 msec default)

- **Inhibition** (see LeabraLayerSpec.cpp for code) – earlier versions of Leabra used an explicit k-Winners-Take-All inhibition function, but the FFFB equations here are much simpler and produce desirable flexibility in overall activation levels:
  - **ffi** = ff * MAX(netin.avg - ff0, 0)
    - feedforward component of inhibition with ff multiplier (1 by default) — has ff0 offset and can't be negative (that's what the MAX(.. ,0) part does).
    - **netin.avg** is average of net variable across unit group or layer, depending on what level this is being computed at (both are supported)
  - **fbi** += fb_dt * (fb * acts.avg - fbi)
    - feedback component of inhibition with fb multiplier (1 by default) — requires time integration to dampen oscillations that otherwise occur — fb_dt = 1/1.4 default
  - **gc_i** = gi * (ffi + fbi)
    - total inhibitory conductance, with global gi multiplier — default of gi=1.8 typically produces good sparse distributed representations in reasonably large layers (25 units or more)

- **Membrane potential** (see LeabraUnitSpec.cpp for code)

- **I_net** = net * (e_rev.e - v_m) + gc_l * (e_rev.l - v_m) + gc_i * (e_rev.i - v_m) + noise
  - net current = sum of individual ionic channels: e = excitatory, l = leak (gc_l is a constant, 0.1 default), and i = inhibitory
  - e_rev are reversal potentials: in normalized values derived from biophysical values, e_rev.e = 1, .l = 0.3, i = 0.25
  - noise is typically gaussian if added
- if ex: **I_net** += g_bar.l * exp_slope * exp((v_m - thr) / exp_slope)
  - this is the exponential component of AdEx, if in use (typically only for discrete spiking), exp_slope = .02 default
- **v_m** += dt.integ * dt.vm_dt * (I_net - adapt)
  - in , we use a simple midpoint method that evaluates v_m with a half-step time constant, and then uses this half-step v_m to compute full step in above I_net equation. vm_dt = 1/3.3 default.
  - v_m is always computed as in discrete spiking, even when using rate code, with v_m reset to vm_r etc — this provides a more natural way to integrate adaptation and short-term plasticity mechanisms, which drive off of the discrete spiking.
- **I_net_r** = net * (e_rev.e - v_m_eq) + gc_l * (e_rev.l - v_m_eq) + gc_i * (e_rev.i - v_m_eq) + noise
  - rate-coded version of I_net, to provide adequate coupling with v_m_eq.
- **v_m_eq** += dt.integ * dt.vm_dt * (I_net_r - adapt)
  - the *equilibrium* version of the membrane potential does *not* reset with spikes, and is important for rate code per below

- **Activation** (see LeabraUnitSpec.cpp for code)
  - **g_e_thr** = (gc_i * (e_rev_i - thr) + gc_l * (e_rev_l - thr) - adapt) / (thr - e_rev.e)
    - the amount of excitatory conductance required to put the neuron exactly at the firing threshold, thr = .5 default.
  - if(v_m > spk_thr) { spike = 1; v_m = vm_r; I_net = 0.0 } else { spike = 0 }

    - spk_thr is spiking threshold (1.2 default, different from rate code thr), vm_r = .3 is the reset value of the membrane potential after spiking — we also have an optional refractory period after spiking, default = 3 cycles, where the vm equations are simply not computed, and vm remains at vm_r.
    - if using spiking mode, then **act** = spike, otherwise, rate code function is below

- `if(v_m_eq <= thr) { new_act = NXX1(v_m_eq - thr) } else { new_act = NXX1(net - g_e_thr) }`
    - it is important that the time to first "spike" be governed by v_m integration dynamics, but after that point, it is essential that activation drive directly from the excitatory conductance (g_e or net) relative to the g_e_thr threshold — activation rates are linear in this term, but not even a well-defined function of v_m_eq — earlier versions of Leabra only used the v_m_eq-based term, and this led to some very strange behavior.
    - NXX1 = noisy-x-over-x+1 function, which is implemented using a lookup table due to the convolving of the XX1 function with a gaussian noise kernel
    - `XX1(x) = gain * x / (gain * x + 1)`
    - gain = 100 default
- **act_nd** `+= dt.integ * dt.vm_dt * (new_act - act_nd)`
    - non-depressed rate code activation is time-integrated using same vm_dt time constant as used in v_m, from the new activation value
- **act** `= act_nd * syn_tr (or just act_nd)`
    - if short-term plasticity is in effect, then syn_tr variable reflects the synaptic transmission efficacy, and this product provides the net signal sent to the receiving neurons. otherwise syn_tr = 1.
- **adapt** `+= dt.integ * (adapt.dt * (vm_gain * (v_m - e_rev.l) - adapt) + spike * spike_gain)`
    - adaptation current — causes rate of activation / spiking to decrease over time, adapt.dt = 1/144, vm_gain = 0.04, spike_gain = .00805 defaults

Learning

Learning is based on running-averages of activation variables, described first:

- **Running averages** computed continuously every cycle, and note the compounding form (see LeabraUnitSpec.cpp for code)

- **avg_ss** += dt.integ * ss_dt * (act_nd - avg_ss)
  - super-short time scale running average, ss_dt = 1/2 default — this was introduced to smooth out discrete spiking signal, but is also useful for rate code
- **avg_s** += dt.integ * act_avg.s_dt * (avg_ss - avg_s)
  - short time scale running average, s_dt = 1/2 default — this represents the "plus phase" or actual outcome signal in comparison to avg_m
- **avg_m** += dt.integ * act_avg.m_dt * (avg_s - avg_m)
  - medium time-scale running average, m_dt = 1/10 average — this represents the "minus phase" or expectation signal in comparison to avg_s
- **avg_l** += avg_l.dt * (avg_l.gain * avg_m - avg_l); avg_l = MAX(avg_l, min)
  - long-term running average — this is computed just once per learning trial, *not every cycle* like the ones above — gain = 2.5 (or 1.5 in some cases works better), min = .2, dt = .1 by default
  - same basic exponential running average as above equations
- **avg_s_eff** = m_in_s * avg_m + (1 - m_in_s) * avg_s
  - mix in some of the medium-term factor into the short-term factor — this is important for ensuring that when neuron turns off in the plus phase (short term), that enough trace of earlier minus-phase activation remains to drive it into the LTD weight decrease region — m_in_s = .1 default.
  - this is now done at the unit level — previously was done at the connection level which is much less efficient!

- *Optional, on by default:* dynamic modulation of amount of Hebbian learning, based on avg_l value and level of err in a given layer — these factors make a small (few percent) but reliable difference in overall performance across various challenging tasks — they can readily be omitted in favor of a fixed avg_l_lrn factor of around 0.0004 (with 0 for target layers — it doesn't make sense to have any Hebbian learning at output layers):
  - **avg_l_lrn** = avg_l.lrn_min + (avg_l - avg_l.min) * ((avg_l.lrn_max - avg_l.lrn_min) / avg_l.gain - avg_l.min))
    - learning strength factor for how much to learn based on avg_l floating threshold — this is dynamically modulated by strength of avg_l itself, and this turns out to be critical — the

amount of this learning increases as units are more consistently active all the time (i.e., "hog" units). avg_l.lrn_min = 0.0001, avg_l.lrn_max = 0.5. Note that this depends on having a clear max to avg_l, which is an advantage of the exponential running-average form above.

- **avg_l_lrn** `*= MAX(1 - cos_diff_avg, 0.01)`
  - also modulate by time-averaged cosine (normalized dot product) between minus and plus phase activation states in given receiving layer (cos_diff_avg), (time constant 100) — if error signals are small in a given layer, then Hebbian learning should also be relatively weak so that it doesn't overpower it — and conversely, layers with higher levels of error signals can handle (and benefit from) more Hebbian learning. The MAX(0.01) factor ensures that there is a minimum level of .01 Hebbian (multiplying the previously-computed factor above). The .01 * .05 factors give an upper-level value of .0005 to use for a fixed constant avg_l_lrn value — just slightly less than this (.0004) seems to work best if not using these adaptive factors.

- **Learning equation** (see LeabraConSpec.h for code) — most of these are intermediate variables used in computing final dwt value
  - **srs** `= ru->avg_s_eff * su->avg_s_eff`
    - short-term sender-receiver co-product — this is the intracellular calcium from NMDA and other channels
  - **srm** `= ru->avg_m * su->avg_m`
    - medium-term sender-receiver co-product — this drives dynamic threshold for error-driven learning
  - **dwt** `+= lrate * [ m_lrn * XCAL(srs, srm) + ru->avg_l_lrn * XCAL(srs, ru->avg_l)]`
    - weight change is sum of two factors: error-driven based on medium-term threshold (srm), and BCM Hebbian based on long-term threshold of the recv unit (ru->avg_l)
    - in earlier versions, the two factors were combined into a single threshold value, using normalized weighting factors — this was more elegant, but by separating the two apart, we allow the hebbian component to use the full range of the XCAL function (as compared to the relatively small avg_l_lrn factor applied *inside* the threshold computation). By multiplying by avg_l_lrn outside the XCAL equation, we get the desired contrast enhancement

property of the XCAL function, where values close to the threshold are pushed either higher (above threshold) or lower (below threshold) most strongly, and values further away are less strongly impacted.

  * m_lrn is a constant and is typically 1.0 when error-driven learning is employed (but can be set to 0 to have a completely Hebbian model).
  * XCAL is the "check mark" linearized BCM-style learning function (see figure) that was derived from the Urakubo Et Al (2008) STDP model, as described in more detail in the CCN textbook: `http://ccnbook.colorado.edu`
  * `XCAL(x, th) = (x < d_thr) ? 0 : (x > th * d_rev) ? (x - th) : (-x * ((1-d_rev)/d_rev))`
  * d_thr = 0.0001, d_rev = 0.1 defaults
  * x ? y : z terminology is C syntax for: if x is true, then y, else z

- **Momentum** — as of version 8.2.0, momentum is turned on by default, and has significant benefits for preventing hog units by driving more rapid specialization and convergence on promising error gradients.

  - **`dwavg`** `= MAX(dwavg_dt_c * dwavg, ABS(dwt))`
    * increment the running-average weight change magnitude (dwavg), using abs (L1 norm) instead of squaring (L2 norm), and with a small amount of decay: dwavg_dt_c = 1 - .001 — software uses dwavg_tau = 1000 as a time-constant of this decay: `dwavg_dt_c = 1 - 1/dwavg_tau`.
  - **`moment`** `= m_dt_c * moment + dwt`
    * increment momentum from new weight change — `m_dt_c = 1 - 1/m_tau` where m_tau = 20 trial time constant for momentum integration by default, which works best (i.e., m_dt_c = .95 — .9 (m_tau = 10) is a traditionally-used momentum value that also works fine but .95 (m_tau = 20) works better for most cases.
  - `if(dwavg != 0) dwt = moment / MAX(dwavg, norm_min); else dwt = moment`
    * set the weight change used by following weight update equation to use momentum, normalized by dwavg if available (nonzero) — this normalization is used in RMSProp, ADAM, and other related algorithms.

- **Weight update equation** (see LeabraConSpec.h for code) (see below for alternative version using differential fast vs. slow weights, not used by default)
  - The **fwt** value here is the linear, non-contrast enhanced version of the weight value, while **wt** is the sigmoidal contrast-enhanced version, which is used for sending netinput to other neurons. One can compute fwt from wt and vice-versa, but numerical errors can accumulate in going back-and forth more than necessary, and it is generally faster to just store these two weight values (and they are needed for the slow vs. fast weights version show below).
  - `dwt *= (dwt > 0) ? (1-fwt) : fwt`
    - soft weight bounding — weight increases exponentially decelerate toward upper bound of 1, and decreases toward lower bound of 0. based on linear, non-contrast enhanced fwt weights.
  - **`fwt`** `+= dwt`
    - increment the linear weights with the bounded dwt term
  - **`wt`** `= SIG(fwt)`
    - new weight value is sigmoidal contrast enhanced version of fast weight
    - `SIG(w) = 1 / (1 + (off * (1-w)/w)^gain)`
  - **`dwt`** `= 0`
    - reset weight changes now that they have been applied.

- *Optional, on by default:* **Weight Balance** — this option attempts to maintain more balanced weights across units, to prevent some units from hogging the representational space, by changing the rates of weight increase and decrease in the soft weight bounding function, as a function of the average receiving weights:
  - `dwt *= (dwt > 0) ? wb_inc * (1-fwt) : wb_dec * fwt`
    - wb_inc = weight increase modulator, and wb_dec = weight decrease modulator (when these are both 1, this is same as standard, and this is the default value of these factors)
  - `wt_avg =`
    - average of all the receiving weights — computed *per projection* (corresponding to a dendritic branch perhaps)
  - `if (wt_avg > hi_thr) then wbi = gain * (wt_avg - hi_thr); wb_inc = 1 - wbi; wb_dec = 1 + wbi`

    - If the average weights are higher than a high threshold (hi_thr = .4 default) then the increase factor wb_inc is reduced, and the decrease factor wb_dec is increased, by a factor wbi that is determined by how far above the threshold the average is. Thus, the higher the weights get, the less quickly they can increase, and the more quickly they decrease, pushing them back into balance.

- `if (wt_avg < lo_thr) then wbd = gain * (wt_avg - lo_thr); wb_inc = 1 - wbd; wb_dec = 1 + wbd`
    - This is the symmetric version for case when weight averages are below a low threshold (lo_thr = .2), and the weight balance factors go in the opposite direction (wbd is negative), causing weight increases to be favored over decreases.
- The hi_thr and lo_thr parameters are specified in terms of a target weight average value `trg = .3` with a threshold `thr=.1` around that target value, with these defaults producing the default .4 and .2 hi and lo thresholds respectively.
- A key feature of this mechanism is that it does not change the sign of any weight changes, including not causing weights to change that are otherwise not changing due to the learning rule. This is not true of an alternative mechanism that has been used in various models, which normalizes the total weight value by subtracting the average. Overall this weight balance mechanism is important for larger networks on harder tasks, where the hogging problem can be a significant problem.

Deep Context

At the end of every plus phase, a new deep-layer context net input is computed from the dot product of the context weights times the sending activations, just as in the standard net input:

$$\eta = \langle x_i w_{ij} \rangle = \frac{1}{n} \sum_i x_i w_{ij} \tag{2}$$

This net input is then added in with the standard net input at each cycle of processing.

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 (Elman, 1990, 1991;

Jordan, 1989; Elman et al., 1996). 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 (e.g., Cleeremans, Servan-Schreiber, & McClelland, 1989; Cleeremans, 1993). 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. 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 23), 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 msec, 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 23).

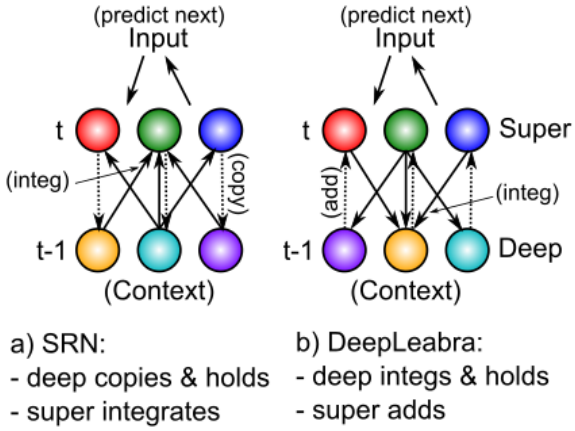

Figure 23: How the DeepLeabra temporal context computation compares to the SRN mathematically. **a)** In a standard SRN, the context (deep layer biologically) is a copy of the hidden activations from the prior time step, and these are held constant while the hidden layer (superficial) units integrate 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.

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 (Thomson, 2010; Thomson & Lamy, 2007; Schubert, Kotter, & Staiger, 2007). 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 msec 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).

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 msec 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 msec, which is well within the time constants in the relevant calcium and second messenger pathways

involved in synaptic plasticity (Urakubo et al., 2008; Bear & Malenka, 1994).

Finally, we note that we had proposed a different, more limited version of this overall DeepLeabra framework previously, which we referred to as *LeabraTI* (temporal integration) (Kachergis, Wyatte, O'Reilly, de Kleijn, & Hommel, 2014). The LeabraTI model hypothesized that higher areas attempt to reconstruct the activation states over the superficial layers of the areas below them, which raised many problems having to do with creating a plausible (and computationally effective) difference between the minus and plus phase states of these areas. Thus, from the perspective of our current framework, the configuration of the TRC neurons within the overall network seems suspiciously ideal for their use as a projection-screen-like substrate for predictive auto-encoder learning. Furthermore, using a single layer driven bidirectionally for the visible layer neurons as we do with the TRC neurons is much more efficient and natural than the two separate layers (input and output) that are required in the typical feedforward SRN framework.