Deep Predictive Learning: A Comprehensive Model of Three Visual Streams

Randall C. O'Reilly, et al
Department of Psychology and Neuroscience
University of Colorado Boulder
345 UCB
Boulder, CO 80309
randy.oreilly@colorado.edu

August 30, 2017

Supported by: ONR grant N00014-13-1-0067, ONR N00014-10-1-0177, ONR D00014-12-C-0638, and Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract number D10PC20021. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained hereon are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

Abstract

How does the neocortex learn? Answers to this fundamental question typically differ across levels, with biologists, computer scientists, and cognitive psychologists each giving different answers. This paper presents a comprehensive framework that spans these levels and leaves no major gaps, providing a potentially complete answer that is directly supported by extensive available data at each level. Broadly, learning is based on making predictions about what the senses will report at 100 msec (alpha frequency) intervals, and adapting synaptic weights to improve the accuracy of such predictions. The pulvinar nucleus of the thalamus serves as a "projection screen" upon which these predictions are generated, through the collaborative inputs of many different brain areas, across multiple levels of abstraction. The otherwise peculiar sparse, strong, and likely non-plastic driving inputs into the pulvinar from layer 5IB (intrinsic bursting) neurons (which fire bursts roughly every 100 msec) provide the target answer, and the temporal difference between the prediction and this answer reverberates throughout the cortex, driving synaptic changes. This form of synaptic plasticity can be derived directly from a highly-accurate biophysical model of STDP (spike-timing dependent plasticity), and builds on the well-validated BCM (Bienenstock, Cooper & Munro, 1982) learning framework. In the domain of vision, these core mechanisms of predictive learning require a carefully-organized developmental progression and anatomical organization of brain areas in order to achieve their full potential, ultimately producing at the highest levels of the ventral *What* pathway in our computational model the essential invariant object-level representations that are otherwise dependent on explicit high-level categorical training signals in most other models. This model shows that compact, high-level abstract representations are essential for accurate prediction of low-level sensory inputs, and that the collective prediction error over the pulvinar can be progressively partitioned, first by a pure abstract dorsal *Where* pathway, then by a *What* * *Where* integration pathway (in V3 and V5/MT), and finally by the ventral *What* pathway. Thus, a full systems-level framework, again supported by extensive data, is required for full realization of this theory, which itself makes a number of testable predictions and is shown here to account for several critical empirical findings.

Introduction

What is the nature of the remarkable neocortical learning mechanisms that result, after many years of experience, in our considerable perceptual and cognitive abilities? 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 cognitive psychological theories. Despite many advances at each of these levels, we still lack an overall framework with the key elements of a comprehensive answer to this question, that has the potential to integrate across these different levels in a mutually compatible way. In this paper, such a framework is presented, providing a broad and deep integration of many different sources of data and theoretical ideas coming from many different researchers. This framework is implemented in a computer model that demonstrates both its computational power and its ability to account for a wide range of data. This model encompasses most of the posterior visual neocortex, including both the dorsal *Where* (and *How*) and ventral *What* pathways, along with a critical *third* visual stream, that serves to integrate information from these other two streams (i.e., a *What * Where* stream).

All three of these streams in our model learn *exclusively* from sequences of visual inputs in V1 (primary visual cortex), via the fundamental (and widely advocated) principle of *predictive learning*: learning to predict what visual input will arise next, given what has come before. Anatomically, the pulvinar nucleus of the thalamus plays the role of the *silver screen of the Cartesian theater of the mind* — it is where the predictions are projected (see Discussion section on implications for consciousness, etc). These predictions are generated every 100 msec (10 hz, alpha rhythm), collaboratively by the entire visual neocortex, and are followed immediately by driving inputs (via layer 5IB intrinsic bursting neurons) from V1 and other cortical areas that reflect the bottom-up *ground truth* training signal. By incorporating key developmental and anatomical constraints on top of a core predictive learning framework based on properties of the pulvinar and deep neocortical layers, the interacting visual pathways learn to represent separable factors (*Where, What, What * Where*) that jointly yield highly accurate and generalizable predictions of subsequent visual inputs. Critically, the *What* pathway of the model develops abstract, invariant object representations without any explicit object category inputs (i.e., in a purely unsupervised manner) — such representations are widely recognized as having great adaptive value to an organism (forming the foundation of much of our semantic understanding of the world), but they typically require extensive hand-labeled categorized training data to develop in other extant models.

Computationally, our framework is a form of *multi-level generative model* (generating predictions), and we found that having high-level abstract representations developing *early* (indeed, before anything else in the case of the dorsal *Where* pathway), and providing strong top-down input to lower areas, is essential for learning an effective predictive model. This contrasts with prevalent approaches that typically learn from the outside-in (i.e., progressively from a bottom-up direction, layer-by-layer). Furthermore, the task of predicting the *future* sensory input avoids many of the problems with widely-explored *auto-encoder* models that learn by reconstructing the *current* input, because, as the saying goes, prediction is difficult, *especially about the future*. These auto-encoders require various constraints to avoid degenerate solutions, and it remains unclear whether such models can function in a purely self-organizing manner. Also, although many other forms of generative models are described in terms of predictive learning, they typically do not actually include the critical *about the future* aspect of predictions, and are instead computationally equivalent to auto-encoders (see the Discussion section for more on this issue). We reserve the term *predictive* here exclusively for its *about the future* sense.

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). 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. We argue that this is just one example of a far more pervasive 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 (to appreciate the importance of these predictive mechanisms, try gently nudging your eyeballs to experience what an unpredictable sensory experience feels like). 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. However, it also represents a notoriously challenging learning problem: how can high-level abstract representations develop prior to the lower-level representations that they build upon? Our model successfully addresses this challenge using a variety of different pragmatic solutions, as we detail below.

Core Mechanisms of Predictive Learning

Predictive learning is an old and widely-explored idea (Elman, 1990, 1991; Jordan, 1989; Schuster & Paliwal, 1997; Hawkins & Blakeslee, 2004; George & Hawkins, 2009), which is also gaining renewed interest in some recent deep neural network models (Lotter, Kreiman, & Cox, 2016). In motor control, the notion of a predictive *forward model* that anticipates the outcomes of actions is well-established (Kawato, Furukawa, & Suzuki, 1987; Jordan & Rumelhart, 1992; Miall & Wolpert, 1996), and the current framework advances the notion that the entire neocortex is a forward model for sensory and motor outcomes. What differentiates our approach is the extent of biological structure that is directly reflected in our model, which provides what we argue is the first complete "no missing (or obviously wrong) links" account for how neocortical learning can solve a full-scale complex task, in a way that directly maps onto considerable neurobiological data about how the different brain areas represented in the model actually function. Specifically, our model provides biologically-sound answers to all of the following essential questions:

- *How do local synaptic signals drive plasticity in a way that produces highly-functional learning in the context of a large complex network of interacting brain areas?* Although the biological data, and locality constraints, appear to favor some variant of a Hebbian learning mechanism, computational models consistently show that this form of learning is incapable of solving real-world problems, and that instead some form of error-driven learning is required. The recent resurgence of interest in backpropagation learning models reinforces the idea that this is the most powerful form of neural learning, and we have long argued that the relevant synaptic mechanisms are readily available to support this form of learning (O'Reilly, 1996; O'Reilly, Hazy, & Herd, 2015; O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012; O'Reilly & Munakata, 2000). Specifically, we argue that known biological mechanisms can readily support learning that is sensitive to a temporal difference in the state of both sending and receiving neurons across the synapse, where this temporal difference reflects an error signal as explained below. We were able to derive this learning rule directly from a highly detailed and well-validated model of spike-timing-dependent-plasticity (STDP) (Urakubo, Honda, Froemke, & Kuroda, 2008). However, all of our previous models have relied upon implausible sources (and timing) of error signals (e.g., explicit category label inputs for object recognition, as in most current deep neural network models) — a critical missing link that is

remedied in the current framework.

- *What is the source of error-driven learning signals?* One of the most appealing features of predictive learning is that the relevant error signals are ubiquitous and "free": these systems learn by comparing what actually happens next versus a prediction generated just prior. In this sense they are effectively *unsupervised* or *self-organizing* learning systems, because they do not require any additional source of learning signals. However, unlike older Hebbian-learning based self-organizing models, predictive learning can leverage the power of error backpropagation to drive learning in a deep hierarchy of areas, in a coordinated fashion, to produce much more powerful results. In short, predictive learning seems to be the best way of accounting for the remarkable ability of babies to just soak up the world by staring endlessly as it moves by (Elman, Bates, Karmiloff-Smith, Johnson, Parisi, & Plunkett, 1996). Also, as noted above, we argue that predictive learning is better than the related, but simpler, goal of *auto-encoding* or re-constructing the current inputs (i.e., by learning a generative model that is capable of regenerating these input patterns). Current deep-neural-network auto-encoder models typically adopt a de-noising framework in order to avoid the network learning a degenerate "mindless copying" solution to the problem: the inputs are presented with noise added, and the network is trained to produce the de-noised version (Bengio, Yao, Alain, & Vincent, 2013b; Valpola, 2014; Rasmus, Berglund, Honkala, Valpola, & Raiko, 2015). By contrast, prediction is sufficiently challenging already, and adding the dynamic, temporal aspect to the problem adds many important dimensions of relevance to the real-world survival of organisms, so we think it is overall a much more likely goal for biological learning. Nevertheless, predictive learning can be viewed as a form of auto-encoding (i.e., a *predictive auto-encoder*) in the sense that it is generating low-level visual representations to match actual inputs, and many lessons from auto-encoder networks should be applicable here as well.

- *How are the prediction and actual outcome separately represented, and how is the timing of the prediction and outcome coordinated & organized?* Predictive learning models immediately raise these important and challenging questions, which fortunately admit to direct experimental testing and falsification. The space of possibilities here is large, but we were able to find a particular solution that fits well with some otherwise rather peculiar features of the biology. Specifically, we hypothesize that the higher-order thalamus (i.e., the *pulvinar*) provides the neural substrate for both the predicted and actual outcome, with alternating phases of prediction and outcome organized within the 100 msec / 10 Hz *alpha* cycle that is characteristic of both thalamic and deep neocortical layer firing (this is an evolution of our earlier proposal; Kachergis, Wyatte, O'Reilly, de Kleijn, & Hommel, 2014; O'Reilly, Wyatte, & Rohrlich, 2014c). Thus, instead of having distinct neural substrates dedicated to representing either predictions or outcomes, we hypothesize a particular economy of shared functionality for this common substrate, which is particularly important in supporting the form of biologically-plausible synaptic-level error-driven learning that we had previously developed (O'Reilly, 1996). Specifically, this form of error-driven learning compares two states of network activation over time: an earlier *minus phase* state representing the network's best guess or expectation, versus a subsequent *plus phase* state reflecting the actual outcome (these terms, and the overall temporal-difference framework, were developed originally in the Boltzmann machine; Ackley, Hinton, & Sejnowski, 1985). Whereas we previously had only general speculations about how these phases were organized over time, and what constituted the actual plus phase signal, the predictive-learning-over-the-pulvinar hypothesis, organized in alternating phases within the alpha cycle, provides concrete, testable predictions that we evaluate below.

  Although the prediction and outcome are encoded over the same pulvinar substrate in our model, we do hypothesize that the superficial (4,2,3) and deep (5,6) layers of the neocortex play distinct roles,

with the superficial layers representing the *current state* of the environment and the ongoing internal "mental" state of the organism, while the deep layers are specifically responsible for *generating the prediction* about what will happen next (via direct projections into the pulvinar). Well-established patterns of neocortical connectivity combine with phasic burst firing properties of a subset of deep-layer neurons to effectively shield the deep layers from direct knowledge of the current state, creating the opportunity to generate a prediction. Effectively, the deep layers of the model are briefly closing their "eyes" so that they can have the challenge of predicting what they will "see" next. This phasic disconnection from the current state is essential for predictive learning (even though it willfully hides information from a large population of neurons, which may seem counter-intuitive), and the remarkable convergence of biological properties supporting this phasic disconnection property in the deep neocortical layers provides strong support for our overall hypothesis.

- *How are error signals represented, and transmitted to drive learning?* Many attempts to map error-driven learning and generative models into the brain hypothesize the presence of neurons that explicitly represent the error signal in their firing. However, as reviewed in detail below, we find the available evidence in support of such neurons in the neocortex or thalamus to be weak and subject to compelling alternative explanations. Thus, we favor the *implicit* temporal-difference version of the error signal in our model as described above. To summarize, the entire network interactively *settles* or converges on both a integrated representation of the current state throughout the superficial layers, and its best prediction of what will happen next in the deep layers projecting to the pulvinar, within the first 75 msec period of the overall 100 msec alpha cycle. The full network of brain areas can thus work together to collaboratively produce the best possible such representations, with individual pyramidal neurons sending standard excitatory signals to other pyramidal neurons, amid a background of dynamic *surround* inhibition. Then, when the plus-phase outcome state is experienced over the last 25 msec of the alpha cycle (driven by burst firing of deep layer 5IB intrinsic bursting neurons that send strong feed-forward driving inputs to pulvinar thalamic relay cells (TRC's), as elaborated below), any differences between this outcome state and the prior prediction state are experienced as ripples of propagating activation-state differences emanating from the pulvinar and penetrating throughout the network. Neurons receiving these projections from the pulvinar, both directly and indirectly, learn locally based on the temporal difference in their activation states across this critical alpha-frequency time-cycle. Mathematically, these temporal differences reflect the same error gradient as computed by the explicit error backpropagation algorithm (O'Reilly, 1996), and we have shown that these error gradients propagated as activation signals through multiple interconnected areas are sufficient to train powerful deep object recognition networks (O'Reilly, Wyatte, Herd, Mingus, & Jilk, 2013; Wyatte, Herd, Mingus, & O'Reilly, 2012b; Wyatte, Curran, & O'Reilly, 2012a).

In recognition of the critical predictive role of deep neocortical layers, and the ability to train deep hierarchical networks, we refer to this 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 modeled only on the superficial layers of the neocortex (O'Reilly et al., 2015; O'Reilly et al., 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, Vezoli, Chameau, Falchier, Quilodran, Huissoud, Lamy, Misery, Giroud, Ullman, Barone, Dehay, Knoblauch, & Kennedy, 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.

Learning a Multi-level Generative Model of the Visual World

To complement this core computational framework, we have finally discovered a systems-scale organization of interconnected areas, and a developmental trajectory thereof, that is capable of fully leveraging these predictive learning mechanisms. Our initial attempts to test the DeepLeabra framework followed the widely-adopted idea of a progressive development of hierarchically-organized neocortical areas, proceeding from the outside-in (Shrager & Johnson, 1996; Bengio et al., 2013b; Valpola, 2014; Rasmus et al., 2015; Hinton & Salakhutdinov, 2006). Specifically, lower-level visual areas such as V1 and V2 develop their representations first, predicting whatever they can at this lower-level, and then higher areas are progressively added to build upon these lower levels and develop higher-level representations *learned from the residual prediction errors left over from the lower areas*. However, we inevitably found that these models never really learned very well (i.e., they could not do a very good job of predicting what was going to happen in the next 100 msec), nor did we find evidence of useful abstract representations developing in higher areas. Eventually we concluded that this approach may be entirely backward — what if the residual error from relatively impoverished lower-level representations is *not* in fact a sound basis for the formation of useful higher-level abstractions?

Instead, we are now convinced that predictive learning must *start* with as much high-level abstract representation as possible, and focus on learning further such representations as quickly as possible thereafter, *because central, compact, abstract representations of things like spatial motion and object properties are essential for successful predictive models*. Without these coherent, central, higher-level representations, the lower-level predictions are doomed to mediocrity — they will learn a vague, muddled and incoherent predictive model, which does not then provide a good basis for developing higher-level abstract representations at a later stage of learning.

High-level abstract representations are essential because they consolidate and concentrate all the learning about e.g., the structure and relationship of different features of an object in one centralized set of representations. These central representations can much more easily maintain this essential information over time, to support consistent, stable predictions about how an object will appear in the next moment. By contrast, lower-level areas such as V1 or V2 are huge and strongly retinotopically organized, such that any given set of neurons only encodes a relatively small portion of the visual world (e.g., around 1 degree of visual angle). Therefore, the encoding of object properties and motion trajectories in such areas must inevitably be highly diffuse and disconnected, with entirely different populations of neurons representing an object at one moment to the next. Such representations provide a poor basis for accurate predictions, given the underlying stability of object properties, and their current motion trajectories, over time (at least over the 100 msec alpha timescale of relevance here).

In short, as with predictive learning, our new model is based on an old, widely-advocated idea (e.g., Rumelhart & McClelland, 1986; Carpenter & Grossberg, 1987; Pollack, 1990; Dayan, Hinton, Neal, & Zemel, 1995; Rao & Ballard, 1999; Hinton & Salakhutdinov, 2006; Friston, 2005, 2010; Bengio, Courville, & Vincent, 2013a; Clark, 2013): a *generative model* of the environment should proceed from abstract, high-level internal representations out to more detailed, specific representations in the periphery — projecting out through the inverse of the feedforward perceptual hierarchy. This idea is easily stated and compelling, but notoriously difficult to achieve in practice, because of the intrinsic interdependencies among all the different levels of representation required. How can abstract, high-level representations develop if the lower, more specific levels are not yet well-formed? How can we develop the abstract

generalization of "cat" when we don't yet know anything about fur, paws, teeth, etc? Indeed, avoiding this kind of catch-22 circularity is exactly what makes the widely-adopted outside-in approach so appealing.

Our model demonstrates that pursuing a pragmatic, opportunistic, and ultimately *emergent* approach to tackling the fundamental circularity of learning a generative model across multiple levels of abstraction makes sense of many disparate properties of the development and function of the visual system, and results in a highly accurate overall predictive learning model.

In overview form, here are the essential strategies and solutions that we have adopted, and their connections to the corresponding biology:

- *First, it is easy to form spatial abstractions, and learn about both externally-generated object motion, and internally-generated saccade motion.* Unlike the formation of invariant object identity abstractions (in the *What* pathway), spatial location (in retinotopic coordinates at least) can be trivially abstracted by simply aggregating across different feature detectors at a given retinotopic location, resulting in an undifferentiated spatial *blob*. These spatial blob representations can drive high-level, central spatial pathways that can learn to predict where a given blob will move next, based on prior history, basic visual motion filters, and efferent copy inputs of saccadic eye movement plans and motor actions. We start our model off by learning these high-level representations, which correspond well with those in area LIP high in the dorsal visual stream, prior to any significant development of any of the rest of the model. These high-level spatial representations then provide strong top-down drive to the lower levels of the model, giving them access to highly accurate spatial prediction signals. This has the highly beneficial effect of partitioning off this spatial aspect of the overall prediction error, thereby concentrating the residual error signals around the remaining problems described next.

  Biologically, there is increasing evidence that this dorsal spatial pathway develops first, and furthermore that there are specific developmental changes in connectivity in relevant areas including the pulvinar, V1, and LIP that specifically support this early development (Bridge, Leopold, & Bourne, 2016). Furthermore, connectivity analyses show that one of the very rare asymmetric pathways in the visual system goes directly from V1 to LIP (Markov, Ercsey-Ravasz, Gomes, R, Lamy, Magrou, Vezoli, Misery, Falchier, Quilodran, Gariel, Sallet, Gamanut, Huissoud, Clavagnier, Giroud, Sappey-Marinier, Barone, Dehay, Toroczkai, Knoblauch, Essen, C, & Kennedy, 2014a).

- *There are **two** residual problems that need to be solved after the pure spatial* Where *problem has been factored out: the traditional* What *problem of representing visual object properties in an invariant manner, and the problem of integrating both* What *and* Where *information for generating highly accurate visual predictions.* Each of these problems presents its own distinct challenges, and each benefits from having its own dedicated hierarchy of neural processing (which nevertheless need to interact extensively with the others). In terms of further partitioning the residual prediction errors, we show that by including this *What * Where* pathway, which we hypothesize may involve area MT (V5) and dorsal prelunate (DP) cortex, the residual error associated with developing a high-quality abstract representation of object features is thereby concentrated in the remaining *What* pathway (areas V4 and TEO) — only then are these representations able to form.

  TODO: Plaut et al papers on visual pathways!

  Biologically, there has been considerable debate about the true extent of separation between the *What* vs. *Where* pathways, and it is evident overall that there is considerable interconnectivity. By positing a third visual stream, whose job it is to integrate *What* and *Where* information, we can potentially make more sense of all this interconnectivity. More generally, the overall objective of learning to accurately predict what will be seen next makes it clear that these areas must interact with

each-other extensively, and our model requires extensive cross-stream connectivity, despite also exhibiting specialization within-streams. Understanding and optimizing this delicate balancing act of specialization and integration required extensive computational experimentation, and may provide an important set of new insights into the functional organization of visual cortex.

- *The* What *visual pathway takes a relatively long time to develop useful abstract representations, so leveraging its benefits requires a later developmental strengthening of top-down connections from this pathway.* We were unable to find a way for this pathway to develop earlier, as it seems to be dependent on successful learning in the other pathways, consistent with the idea that the successful learning in these other pathways concentrates the residual error on object feature information. Thus, these abstractions are not available early-on to support predictive learning, in violation of the principle that abstract high-level representations are essential. We therefore need to posit a later developmental strengthening of these top-down connections, once the representations have sufficiently developed, and we show that this then results in a significant boost in overall prediction accuracy, which still requires a relatively long time period to fully develop. Biologically, there is various evidence for delayed development of the *What* pathway (Rodman, 1994; Nishimura, Scherf, & Behrmann, 2009).

- *The pulvinar (as a kind of projection screen) broadcasts the main prediction error signal throughout the* What * Where *and* What *streams, and structured interconnections among areas then result in the partitioning of the residual errors to develop specialized pathways.* We have consistently found that our model depends critically on all areas at all levels receiving the main predictive error signal generated by the V1 layer 5IB driver inputs to the pulvinar in the plus phase. This was initially quite surprising at a computational level, as it goes strongly against the classic hierarchical organization of visual processing, where higher areas form representations on top of foundations built in lower areas — how can high-level abstractions be shaped by this very lowest level of error signals? We now understand that the overall learning dynamic in our model is analogous to a *multiple regression*, where each pathway in the model learns by essentially absorbing a component of the overall error signal, such that the residual is then available to drive learning in another pathway. Thus, each factor in this regression benefits by directly receiving the overall error signal, and the process of partitioning out the residuals across brain areas requires specific patterns of interconnectivity that we have managed to discover through a long process of experimentation, guided by various principles that emerged in this process (as detailed below). One clear such principle is that although the pulvinar connections have this unusual flat connectivity pattern, most other connections obey the standard hierarchical pattern of connectivity, and this is essential for supporting the development of increasingly abstract representations in higher areas. Also, this wide broadcast from the pulvinar helps to coordinate all of the different layers and share the emerging prediction directly among them, which likely has important benefits as well.

  Biologically, the pulvinar does indeed interconnect widely with all areas of the cortex, and there is strong evidence for the idea that the lowest-level V1-driven signal drives all the major areas in the *What * Where* and *What* pathways (Shipp, 2003; Kaas & Lyon, 2007). In particular, individual V1 5IB driver neurons have multiple (3-5 or so) strong driving synapses into the pulivinar at different levels, whereas other areas only seem to have a single such driving synapse (Rockland, 1998b, 1996).

In summary, we offer a complex, emergent, yet principled account for how the seemingly intractable problem of simultaneously learning concrete and abstract representations across multiple interconnected areas of the visual cortex can be solved. The working computational model is essential here to demonstrate the success of this approach, given the complex and emergent nature of the learning process. In the

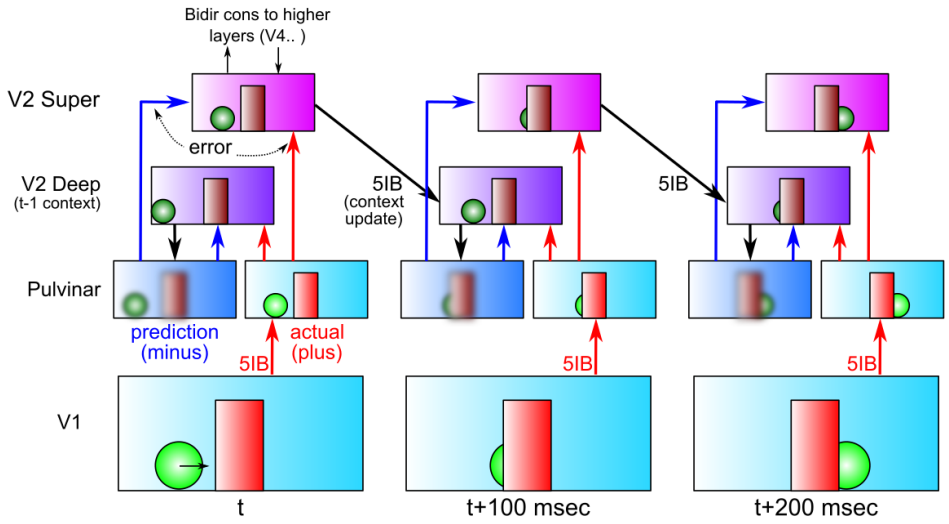

Figure 1: The temporal evolution of information flow in a DeepLeabra model predicting visual sequences, over a period of three alpha cycles of 100 msec each. The Deep network uses the prior 100 msec of context information to generate a prediction or expectation (minus phase) over the pulvinar TRC units 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).

remainder of the paper, we present the model, the simple dynamic environment on which it is trained, and the way in which learning evolves over time in the model. We then use a variety of techniques to probe the nature of what is learned, and the forces that shape this learning, corroborating the overall account just given. Next, we explore the relevant biological data and provide detailed simulations of particularly relevant neural recording data. Because this model simulates such a large portion of the posterior neocortex, the scope of potentially-relevant data is vast, so our treatment is necessarily selective and opportunistic — subsequent work will go into further details. Each of these explorations includes a number of testable predictions from our model, and more general such predictions are outlined in the General Discussion section.

The DeepLeabra Predictive Learning Framework

We begin with an overview of how DeepLeabra models the cortical and thalamic pathways to do predictive learning, in terms of differential functional roles for superficial and deep layers of the neocortex, and loops through the thalamus, and the temporal dynamics of information flow through this circuit.

Figure 1 provides an overall schematic for how predictive auto-encoder learning takes place in our framework, in terms of area V2 predicting the next pattern of activation on V1, over the period of three alpha-cycle "movie frames" (interestingly, actual film-based movies have a frame rate of 24 Hz, which is just over the 2x nyquist sampling limit for a 10 Hz process). 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 to create better representations for making a more accurate prediction next time around.

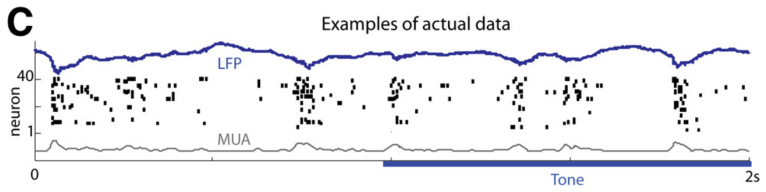

Figure 2: Discretization of a continuous tone stimulus in deep layer neurons at the alpha frequency (Figure 1C from Luczak et al, 2013). This is consistent with the discretized alpha-frequency updating of temporal context representations in the deep layers in the DeepLeabra model. The overall temporal structure was the same for tone on / off conditions, but tone onset did tend to re-synchronize the phase — this kind of stimulus-driven phase resetting and entrainment has not yet been included in our model.

More specifically, the central hypotheses of our model are:

- The neocortex is composed of two separable but tightly interacting sub-networks, superficial and deep / thalamic (pulvinar). The superficial-layer network consists of neocortical layers 4, 2, and 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 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., 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 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 et al., 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. This does not imply that the network converges fully to a stable settled attractor state — just that it 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 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). As such, it only phasically receives new bottom-up input about the state of the environment, triggered by alpha-frequency bursting of the layer 5IB neurons (which is also entrained via thalamocortical networks via various mechanisms (Lorincz et al., 2009; Franceschetti et al., 1995; Saalmann et al., 2012)). 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 (Figure 2; 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), and top-down-specific synchronization (von Stein, Chiang, & König, 2000; van Kerkoerle et al., 2014). 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. We discuss this and other relevant issues in more detail below.

- Biologically, there are two different types of cortical connections into pulvinar TRC neurons (Sherman & Guillery, 2006): strong, sparse *driver* connections originating from 5IB neurons (originally labeled R or type-2; Rockland, 1998a, 1996), and weaker but much more numerous *modulatory* connections originating from 6CT neurons (E or type-1). We depart from the modulatory notion of Sherman and Guillery (2006), and argue that these weaker 6CT inputs are capable of driving TRC activation by themselves, in the form of the minus-phase prediction representation. Indeed, extensive *in vivo* electrophysiological recording data shows constant steady activation of pulvinar neurons across multiple alpha trials worth of time, suggesting that these top-down projections are capable of driving TRC activation in between the 5IB bursting (Bender, 1982; Petersen, Robinson, & Keys, 1985; Bender & Youakim, 2001; Robinson, 1993; Saalmann et al., 2012; Komura, Nikkuni, Hirashima, Uetake, & Miyamoto, 2013). This minus phase is then followed by the strongly-driven 5IB plus-phase representation, which is essentially a copy of the sending layer activations (e.g., V1). To generate the predicted minus-phase state, the layer 6CT neurons rely on integrated inputs from earlier 6 corticocortical (6CC) neurons and 5IB neurons, along with various other largely top-down inputs.

- 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). 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. Indeed, there are well-established capacity constraints of around 2-4 objects (or "fingers of instantiation," FINST's; Pylyshyn, 1989) that can be tracked at any given time, including during the predictive remapping process (Cavanagh et al., 2010). We are generally surprisingly unaware of how much we are *not* tracking, because typically we can just re-access the environment to encode any element we might have initially overlooked (Simons & Rensink, 2005).

Computationally, we show below that these deep / thalamic circuits produce attentional effects consistent with the abstract Reynolds and Heeger (2009) model, while the contributions of the deep layer networks to this function are broadly consistent with the folded-feedback model (Grossberg, 1999). Biologically, the layer 6CT neurons are known to exhibit a multiplicative influence over firing of superficial-layer neurons, in a manner consistent with the Reynolds and Heeger (2009) model (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). We briefly elaborate on these ideas toward the end of this paper, and a subsequent paper will explore them in greater depth.

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 3), and then learned synaptic weights integrate that copied context information into the next

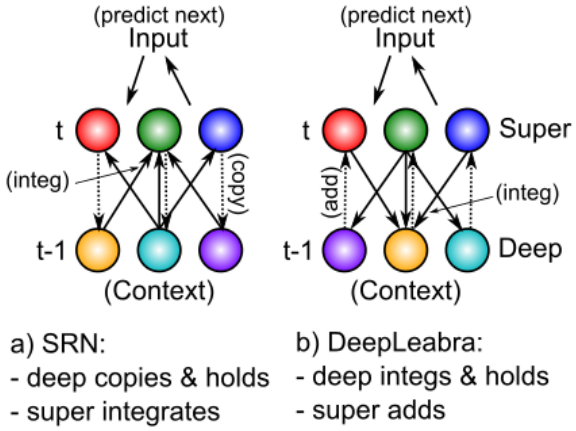

Figure 3: How the DeepLeabra temporal context computation compares to the SRN mathematically. **a)** In a standard SRN, the context (deep layer biologically) is a copy of the hidden activations from the prior time step, and these are held constant while the hidden layer (superficial) units integrate 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.

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 3). 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 et al., 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.

A Comprehensive Model of Three Visual Streams

The DeepLeabra predictive auto-encoder learning mechanisms provide the core engine of our systems-level model of how the three different visual pathways (*Where*, *What * Where*, and *What*) work together to produce highly accurate visual predictions. As summarized earlier, this model requires considerable additional structure and developmental organization to achieve fully successful learning, based on abstract high-level representations driving top-down inputs to the lower areas where the more detailed visual prediction is rendered. The measure of success in this model is not just that it accurately predict the next sensory inputs, but, more importantly, that it develop these high-level abstract representations 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.

The strong correspondence between the specific computationally-motivated network properties and the known biology, reviewed in greater detail here, supports the idea that this model accurately describes how the actual mammalian visual neocortex learns. 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.

The model, which we refer to as the *What-Where-Integration* or *WWI* model, is shown in Figure 4, 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** V1 superficial layer activations, reflecting basic feature extraction (e.g., gabor oriented edge filtering) on retinal input signals, the saccade-related signals 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

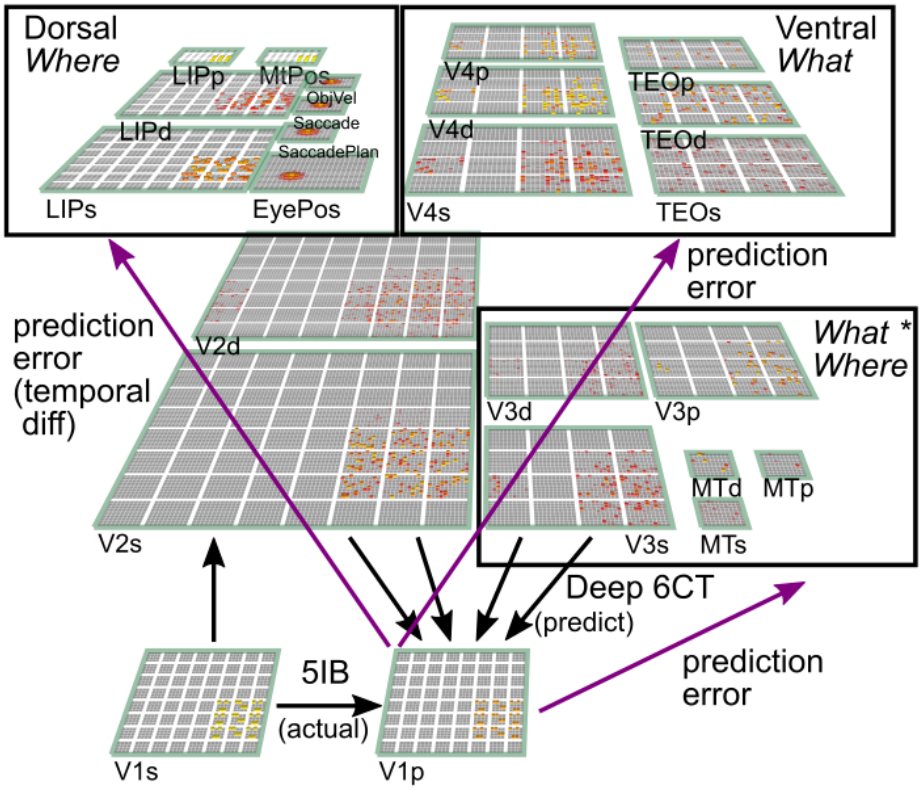

Figure 4: The three-visual-stream deep predictive learning model (What-Where-Integration or WWI model). The dorsal *Where* pathway learns first, using abstracted *spatial blob* representations, to predict where an object will move next, based on prior motion history, visual motion, and saccade efferent 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) "projection screen" 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).

without any external constraints or presumptions. We also have a number of *decoder layers* (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 decoder layers 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. 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 (i.e., the "silver screen of the Cartesian Theater"). Our primary predictive learning error measure is 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

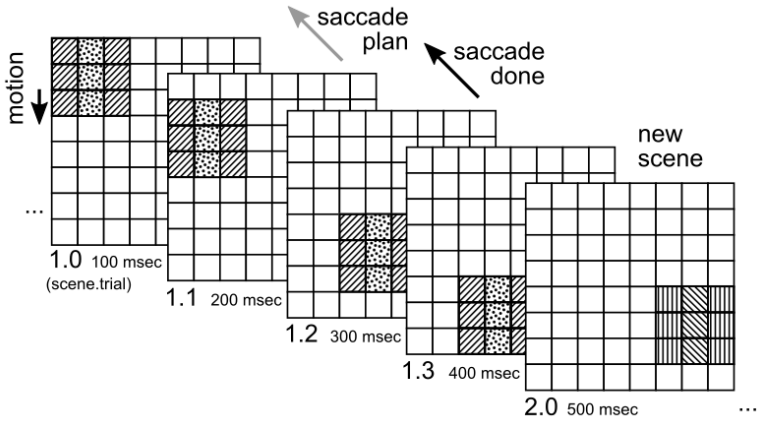

Figure 5: Dynamic visual environment, with 100 different objects composed of two independent sets of features (central column vs lateral flankers, 10 different patterns each), that have a constant motion vector (including the 0,0 no motion case) — a 1 cell per trial downward motion is shown. New scenes 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).

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 et al., 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 (convolution). 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 5) 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 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. To keep things as simple and small as possible, we used an 8x8 grid of V1 hypercolumns (each hypercolumn having 4x4=16 feature bits), with an individual object subtending a 3x3 contiguous grid 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. 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). We use 0-based indexes for referring to trials by number. The object features remain consistent during a given scene, and change randomly for the next scene. Thus, the first trial is unpredictable, and only on the second trial does the network have the ability to make an accurate prediction. For this reason, the predictive learning framework in general requires at least 2 trials of processing for a novel visual input — in combination with our hypothesis of alpha frequency predictive trials, we strongly predict that fixation durations should last at least 200 msec, which appears to be consistent with available data as reviewed below. 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 is 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.

Model Mechanisms

The model uses standard *Leabra* equations (O'Reilly et al., 2015; 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). Earlier versions of Leabra used an explicit k-Winners-Take-All inhibition function, but the FFFB equations (see the Appendix) are much simpler and produce desirable flexibility in overall activation levels. 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

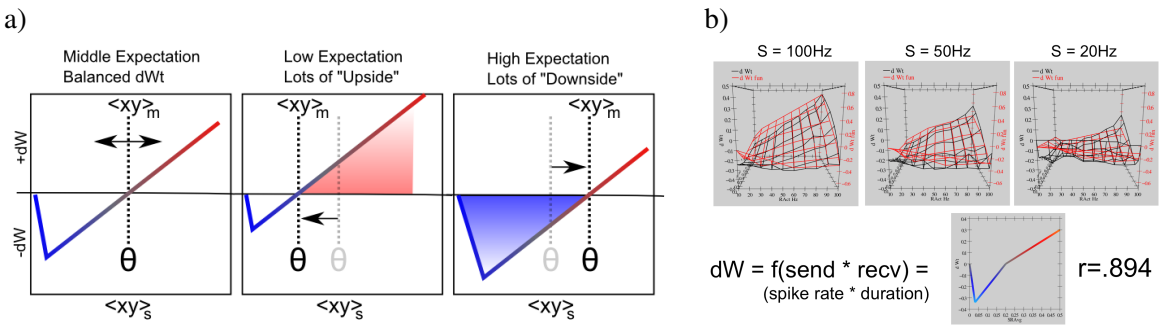

Figure 6: Error-driven synaptic plasticity in Leabra, using the *XCAL* function that is a linearized version of the BCM plasticity function, as derived from the Urakubo et al (2008) STDP model shown in panel (b). a) The threshold θ between weight decrease (-dW, LTD) and weight increase (+dW, LTP) can 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 (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).

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 et al., 2008; Figure 6). 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 6). Interestingly, recent data is consistent with such rapidly adapting thresholds (Lim, McKee, Woloszyn, Amit, Freedman, Sheinberg, & Brunel, 2015). 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 mutually-interconnected areas can build up mutually reinforcing weights, causing the 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 4). 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 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 4) 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

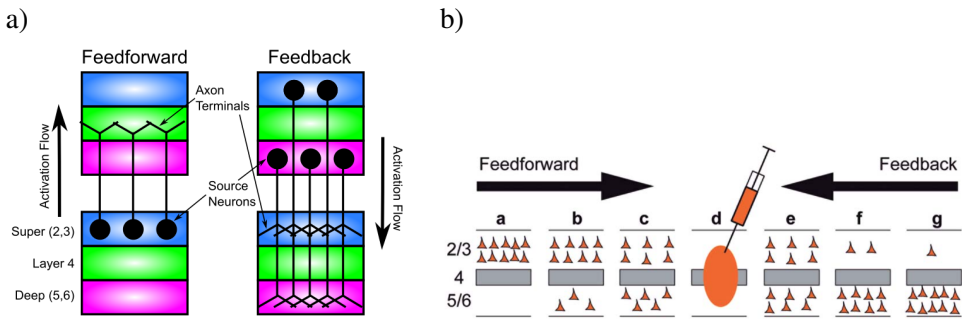

Figure 7: Standard patterns of feedforward and feedback connectivity in neocortex. a) Most feedforward connections originate in superficial layers of lower area, and terminate in layer 4 of higher area. Feedback connections can originate in either superficial or deep layers, and in both cases terminate 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. Overall, these patterns are critical for the functioning of the predictive learning model as explained in the text.

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 — if adding a given connection made little overall difference, or made performance worse, it was left out of the default model. Reassuringly, 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 7 (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). Note that although Markov et al. (2014b) present evidence that the feedforward and feedback pathways in the superficial layers may be supported by separate populations of neurons (in layer 2 vs. 3B), both of these populations receive the same feedforward (via layer 4) and feedback (via layer 1 dendritic tufts) projections, so this may just be more of a wiring difference without strong functional implications — we will explore these issues in later versions of our model.

As noted earlier, it is essential in the DeepLeabra model that the feedforward connections do *not* project

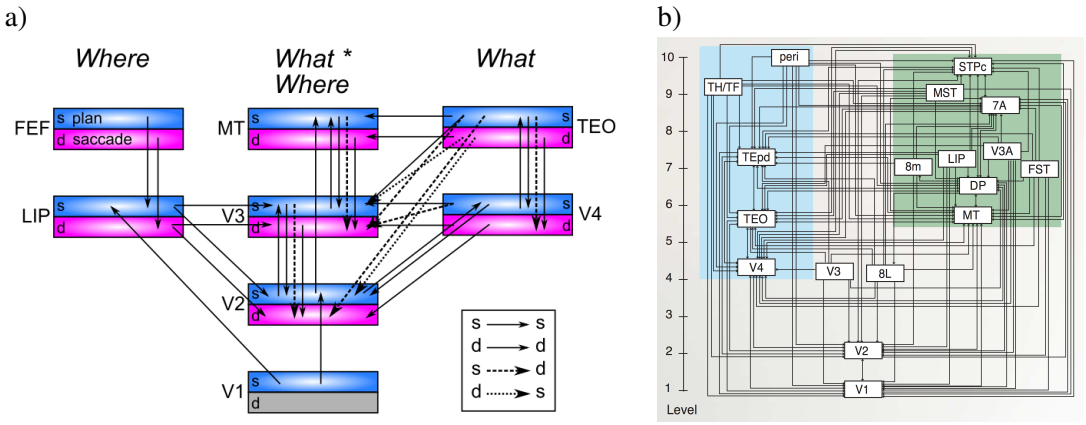

Figure 8: a) Superficial and deep-layer connectivity in the model. Note the repeating motif between hierarchically-adjacent areas, with bidirectional connectivity between superficial layers, and feedback into deep layers from both higher-level superficial and deep layers, according to canonical pattern 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 more pure *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.

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 in the first place. This would be analogous to a short-circuit in electrical terms. Furthermore, as we demonstrate below, it is very important that the feedback connections from superficial layers *do* drive the deep layers directly — 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. More generally, it seems that the deep layers in our model only benefited from top-down projections, not bottom-up ones (which could only be from other deep layers, due to the short-circuit problem). 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 compact, high-level representations (as discussed earlier).

Figure 8 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 outside of the pure *Where* pathway, the characteristic pattern shown in Figure 7 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 V1 is strictly an input layer in this model, so all top-down and deep-layer connectivity was omitted). 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 are essential for allowing more abstract, high-level TEO representations to positively influence the low-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, while VP2 (in lateral pulvinar) does not. We are excited to explore possible contributions of collicular inputs in future models — they may serve as another source of plus-phase training signals, 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 pure 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.

All of these pulvinar maps have a third dimension of organization beyond their 2D retinotopic maps, the *axis of iso-representation* (AIR), which *roughly* reflects the corresponding cortical hierarchy (although it is inverted relative to cortex in the caudal-rostral dimension). The lower visual areas, V1, V2, and V3, project extensively across the AIR dimension (with the densest projections in the most rostral region), and V1 in particular has multiple branches of driving (R, type-2) projections along this dimension (Rockland, 1998b, 1996; Sherman & Guillery, 2006). By contrast, higher areas send these driving projections more caudally along the AIR dimension (corresponding to higher-level areas), while also sending weaker (E, type-1) projections to the more rostral, lower areas. Overall, the connectivity from pulvinar to cortex tends to be reciprocal (symmetric) to the connectivity from cortex to pulvinar.

Our overall 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 so attractive as a substrate for projecting representations on to. 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

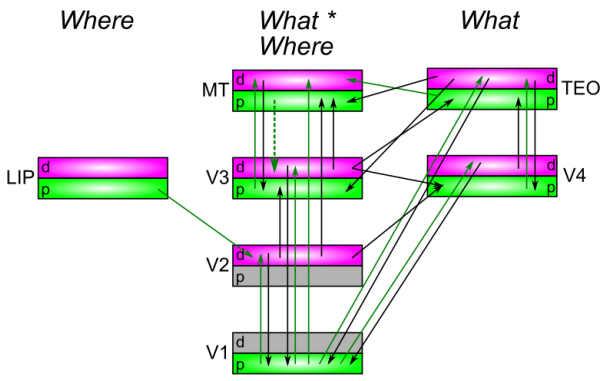

Figure 9: Connectivity for deep layers and pulvinar in the model, which generally mirror the corticocortical pathways (previous figure). Each pulvinar layer (p) receives 5IB driving inputs from the labeled layer (e.g., V1p receives 5IB drivers from V1). In reality these neurons are more distributed throughout 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.

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, 1998b, 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 is inaccurate (but a useful first approximation) — it is really a number of different screens at various levels of abstraction.

Figure 9 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 8) — 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 areas. 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 it *first*, before anything else of significance is attempted, to absorb as much of the predictive error associated with the purely spatial aspect, and thereby drive other areas to take on the remaining *What * Where* and *What* components. 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. 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

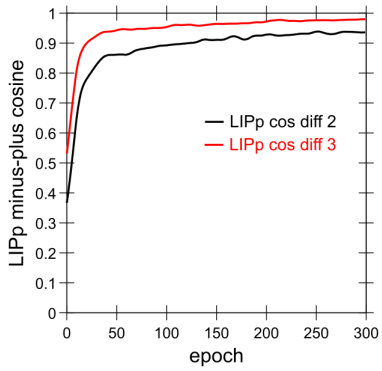

Figure 10: Learning curves for LIP spatial prediction accuracy, measured as cosine between minus and plus phase representations over the LIPp pulvinar layer (perfect accuracy is 1.0). Trial 2 (LIPp cos diff 2, which is the 3rd trial of the sequence) is right after the saccade and thus requires integrating 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 *What* 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.

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 pure *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 a purely 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 4). A later model could explore a more realistic developmental transition of a common MT area, potentially revealing interesting benefits from the early developmental phase.

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 pure *Where* pathway are shown in Figure 10, 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 suspiciously 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 from the start. 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 relatively later in development and continue to develop over a relatively long timecourse (Nishimura et al., 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 various tests, manipulations, and analyses that show 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, which are documented first. The subsequent results section then explores how the model accounts for some detailed empirical data of particular relevance.

The learning curves for the full intact model are shown in Figure 11, showing that the model achieves high levels of predictive accuracy in terms of the cosine difference between the minus and plus phase activation 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 develops a much more systematic, generalizable representation of objects compared to other layers. This is evident in the ability of the *decoder* layer (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=.01 for both). The decoding of TEO is roughly 2x better (i.e., a 2x reduction in error) compared to the V3 layer. Numerically, this is particularly 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 dimensions.

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 of this is developing 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 11a 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

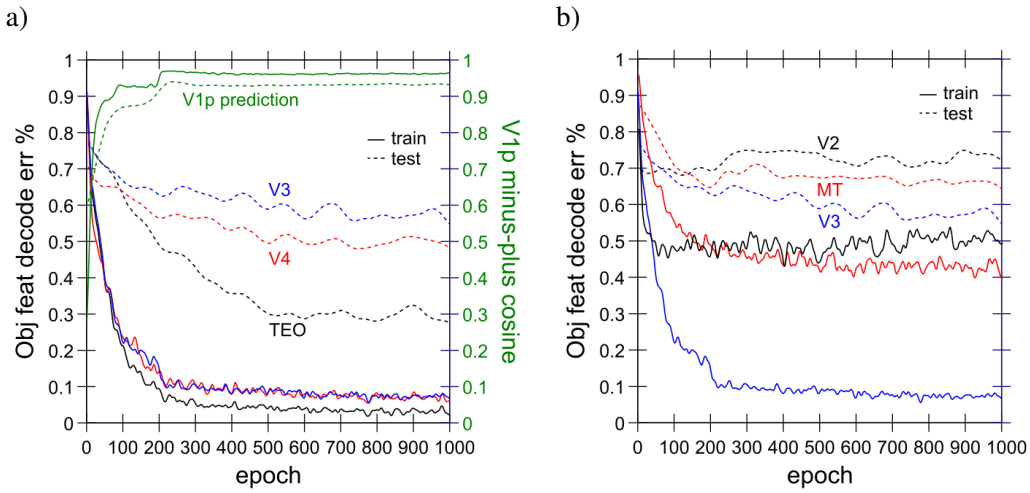

Figure 11: a) Learning curves for full model, showing accuracy (proportion error) in decoding the object features from each of 3 different layers (V3, V4, TEO), and overall prediction accuracy in terms of minus vs. plus phase cosine over the V1p pulvinar layer, at trial 3 (the last trial), which is nearly perfect. Note the discrete jump in prediction accuracy when we turn on the top-down weights from TEO, at epoch 200. The decoding shows a roughly 2x reduction in error for TEO vs. V3, and is especially evident in raw terms for the 10 novel untrained testing items. This shows that TEO has developed much more systematic object representations than those in 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 lower 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 (and computationally, it takes 12 hours to run using 64 processors in parallel on our cluster). Due to the time required, results are from single runs, but we did run multiple replications of several key conditions and they were very reliable.

biology and recent deep neural network models. V2 has very low object decoding accuracy, so V3 produces large gains in object decoding accuracy, but mainly for the trained items — the novel test items show only a modest improvement. Thus, the trained-object decoding accuracy measure does not necessarily indicate that V3 has invariant or compact object representations — just that the information can be extracted by the decoder in any way (albeit within the constraints of a single-layer set of weights). The test-object decoding performance is really the best measure of how systematic and invariant the object representations are, as is evident in the direct analysis of the representations shown next.

Interestingly, the MT layer shows *worse* object decoding accuracy compared to its input layer, V3 (Figure 11b), 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.

We also 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 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

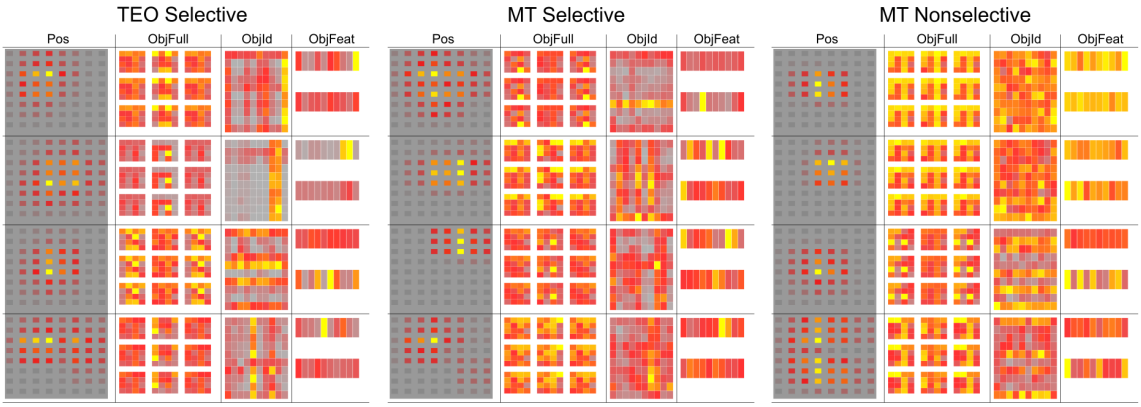

Figure 12: Activation-based receptive fields for TEO vs. MT (superficial layers), selected for relative feature selectivity, and MT non-selective cases. Each cell shows weighted average activation across position and object decoding patterns as a function of unit activity, for 4 target units from each layer (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., the 2nd TEO unit from top 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.

		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.

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. This more complex developmental dynamic will be explored in subsequent work.

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 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 12 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 12) 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. This is overall consistent with available data on TEO neurons, which also encode spatial location along with many other properties (Majaj et al., 2015), and with the general notions of coarse-coded distributed representations of high-dimensional data which are useful for binding (Hinton, McClelland, & Rumelhart, 1986; O'Reilly & Busby, 2002; O'Reilly, Busby, & Soto, 2003; Cer & O'Reilly, 2006), also known as *mixed selectivity* (Fusi, Miller, & Rigotti, 2016).

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 et al., 2015).

Importance of a Deep Hierarchy: Testing Flatter Models

Figure 13a 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

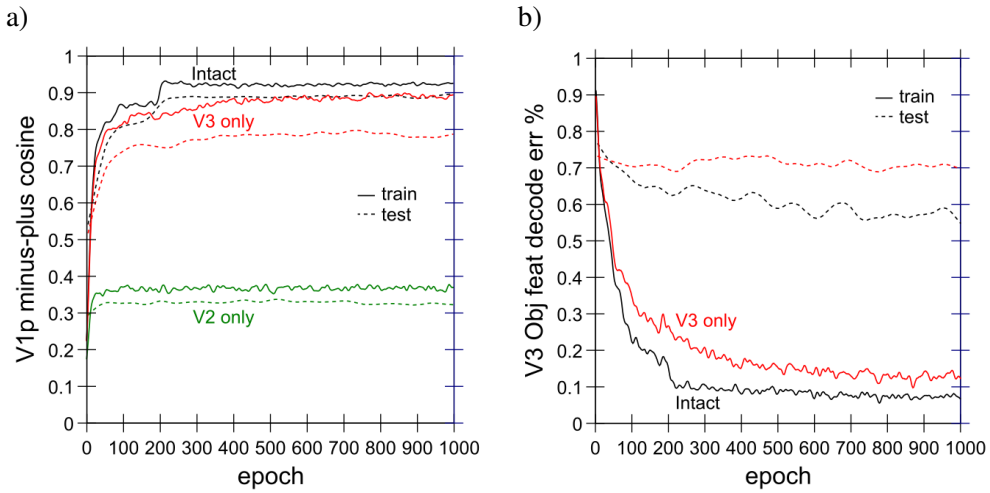

Figure 13: a) Prediction accuracy (minus vs. plus phase cosine over the V1p pulvinar layer), at trial 2 (the post-saccade trial) for model with only V2 (no V3, MT, V4, TEO) or only V3, compared to the full Intact model. A single layer alone (V2 only) cannot do very well, despite getting nearly-perfect 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.

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 the ability to more systematically represent objects. Figure 13b 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 14 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. Here you can also more clearly see (due to the use of a more restricted vertical range in the graph) the significant bump in prediction accuracy in the Intact model right after the top-down connections from TEO are turned on at epoch 200, reflecting the hypothesized delay in maturation of these projections. Interestingly, 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 14b, showing

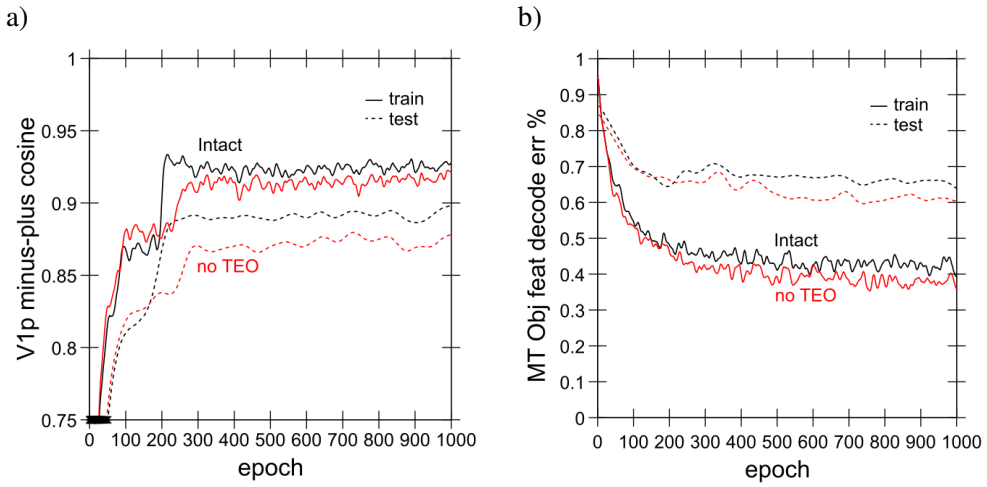

Figure 14: a) Prediction accuracy (as in prior figure) for Intact versus model with no TEO area, showing small but reliable impairment, more for test than trained objects. b) Object feature decoding accuracy from layer MT for Intact vs. no TEO model, showing *improvement* in object detection in MT when TEO is lesioned, consistent with opponent relationship between these pathways.

the 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 (see Appendix for details), 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 15a. 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 10). 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. With larger more complex spatial and object representational spaces in the real system, these effects would likely be magnified considerably.

Figure 15b 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 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.

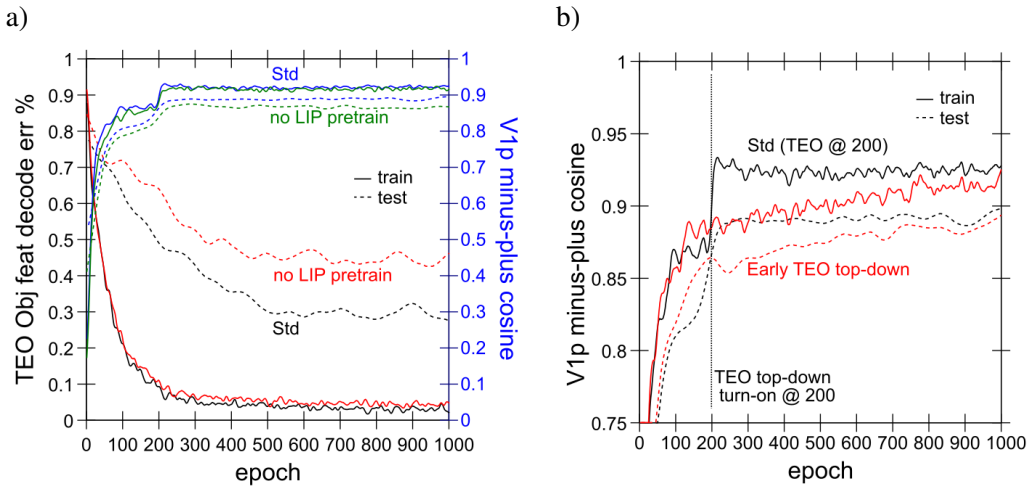

Figure 15: a) Learning without first pretraining the *Where* LIP pathway, which has a significant impact on the development of systematic TEO object representations, particularly for the testing items. This has corresponding effects on V1p prediction accuracy (top lines), again particularly on the testing 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.

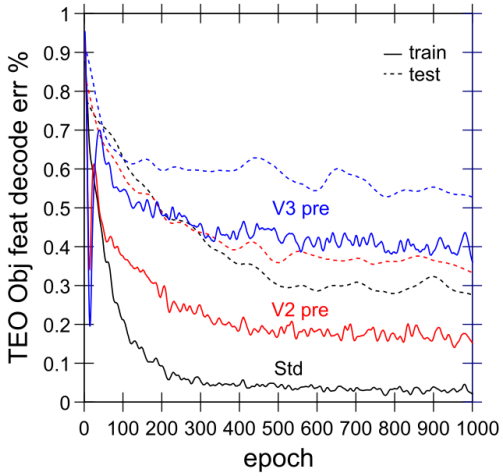

Figure 16: Effects of pretraining using weights from V2 only or V3 only model on object decoding accuracy from the TEO area, as a test of the standard outside-in developmental training approach. This significantly impairs the development of systematic invariant object representations in TEO, presumably by interfering with the prediction error partitioning process, and the top-down influence of more abstract object representations during learning.

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 16 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

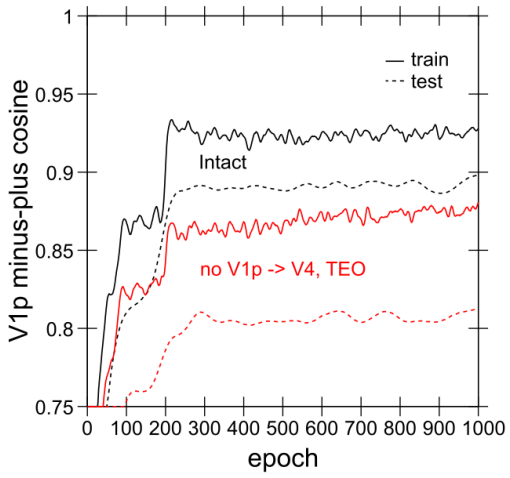

Figure 17: Effects of removing the V1p to V4,TEO projection on overall V1p prediction accuracy, showing similar effects to a TEO lesion, indicating that the *What* pathway is essentially non-functional without this V1p pulvinar projection. Consistent with this, object decoding accuracy in TEO was also completely abolished (not shown).

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 13). 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 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 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 17). However, there are only relatively minimal

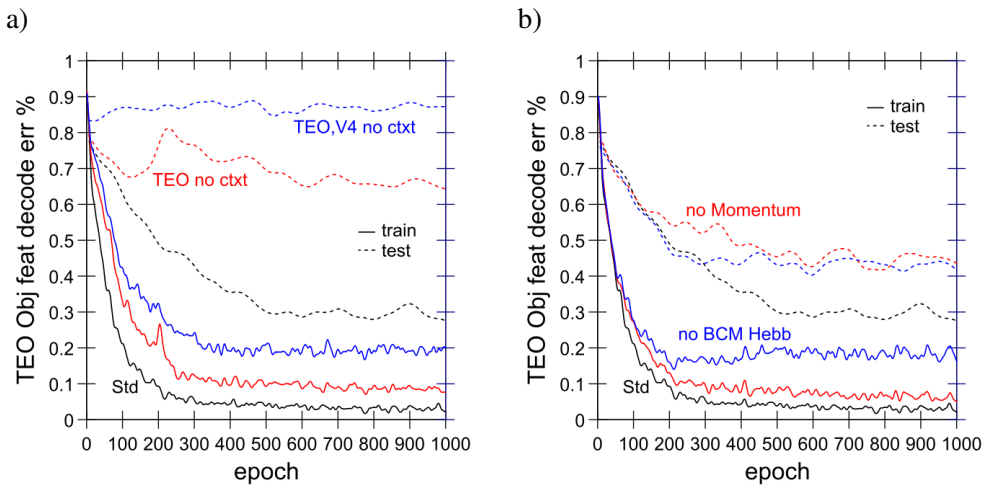

Figure 18: a) Effects of removing the deep-layer context inputs into TEO and V4 + TEO together — this has a major impact on ability to decode object features from TEO, particularly in the case of the novel testing items. b) Effects of not using momentum or BCM-like Hebbian learning.

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 18 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 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 it 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. Future work will explore many other such phenomena.

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), 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 19a).

Figure 19b 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). 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. Instead, it makes much 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 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

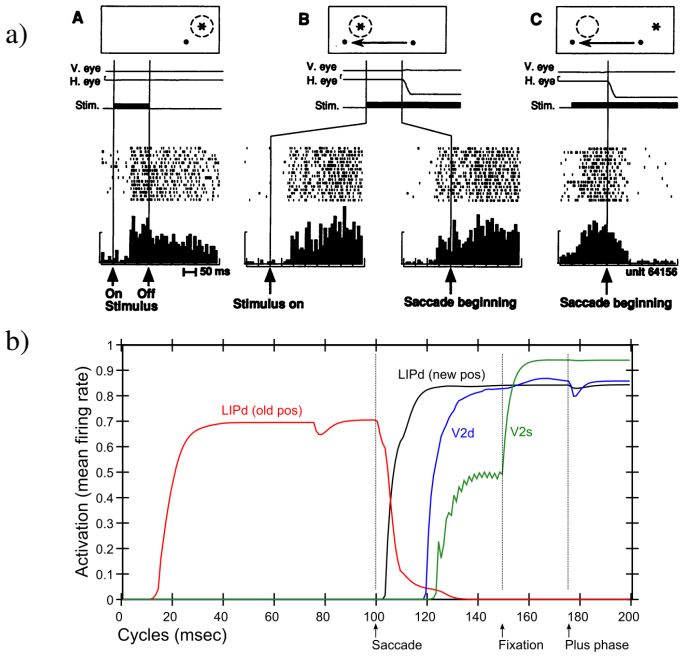

Figure 19: a) Original remapping data in LIP from Duhamel et al (1992). A) shows stimulus (star) response within receptive field (dashed circle) relative to fixation dot (upper right of fixation). B) Just prior to monkey making a saccade to new fixation (moving left), stimulus is turned on in receptive 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.

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).

In a future, larger-scale model, we plan to address the potentially important differences between microsaccades (less than 1 degree) and full saccades (Martinez-Conde, Otero-Millan, & Macknik, 2013; Martinez-Conde, Macknik, & Hubel, 2004). Unlike full saccades, microsaccades *can* be predicted within the typical receptive field sizes of V2 neurons, and there is evidence that visual motion signals are also used to predict the outcome of such saccades (along with passive visual drift which is also prevalent at these small scales). Interestingly, the new cortical hierarchy analysis by Markov et al. (2014b) (Figure 8b) separates the frontal eye fields (FEF, area 8) into two parts, at different locations in the hierarchy. The part of FEF responsible for small-displacement saccades (8L) is located at the same level as V3, while the large-displacement part (8m) is higher up at the level of LIP and is assumed to provide the saccade signals in our current model. Thus, this microsaccade system involving 8L, V3, and V2 may provide a rich source of additional predictive learning training for shaping these high-resolution, lower areas of the visual system.

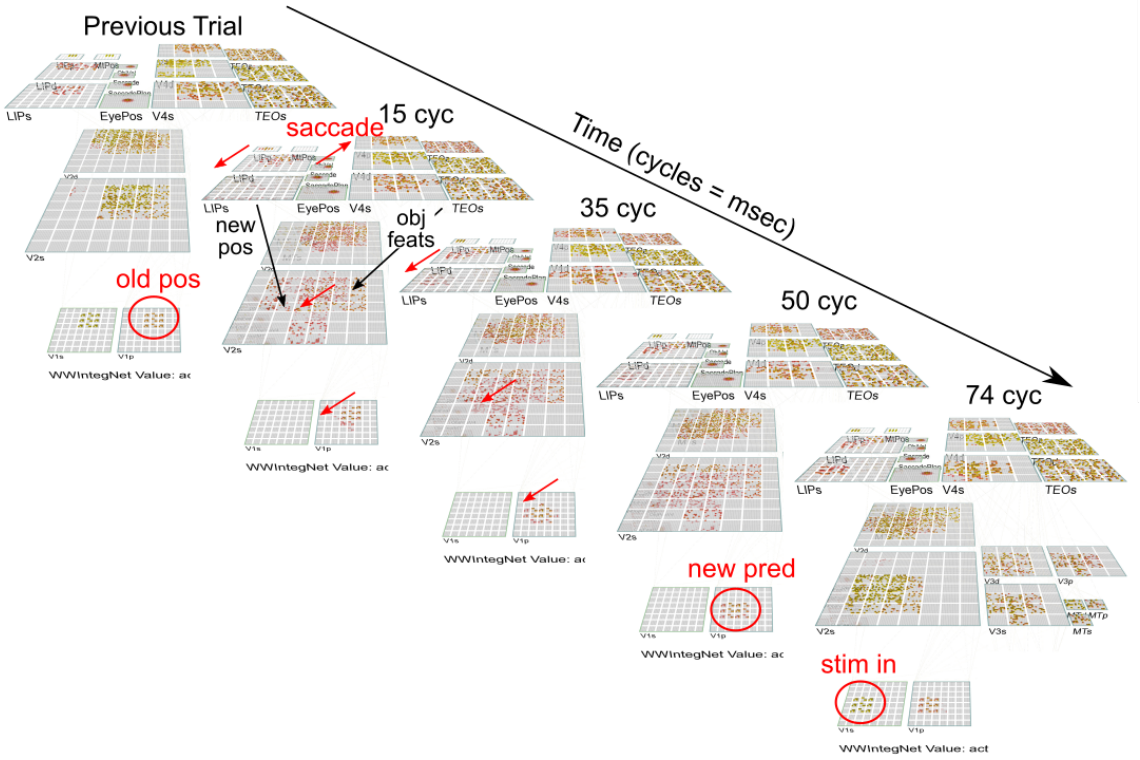

Figure 20: Predictive remapping in the entire model, showing the early movement of LIPd activation that then drives LIPp (which then updates LIPs as well), and sends top-down input to V3 (occluded) and V2. These lower layers receive top-down input from TEO providing a representation of object features, 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.

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 21). 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 of similar such effects. Because of the simplicity of our visual environment, we could not directly replicate the existing experiments, 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 21 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.

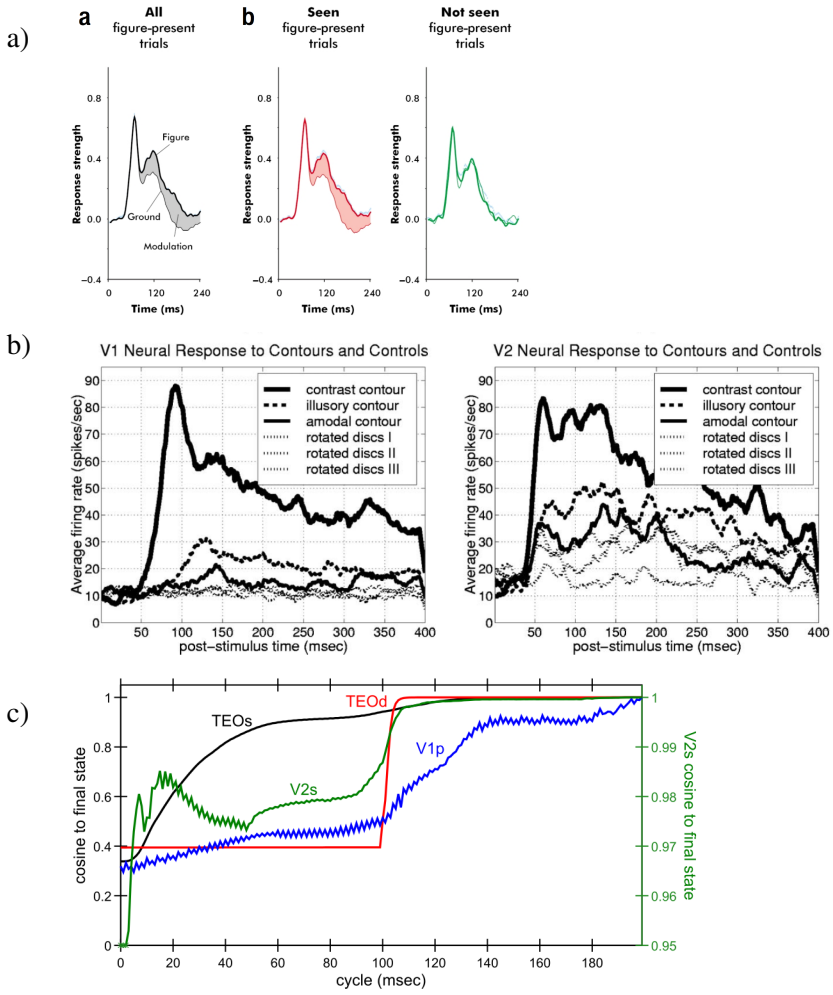

Figure 21: Top-down effects on lower-level neural firing. a) Top-down modulation of V1 firing as a function of a texture-defined figure/ground stimulus, emerging after 100 msec (one alpha trail) in monkeys, specifically as a function of whether the monkey makes a behavioral response indicating that the 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 object features (see supplemental information for a video of actual network states). 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.

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). 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 21a,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-lasting 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.

Instead, we investigated another possible mechanism behind relatively higher activation levels for unpredicted outcomes, that might help to explain why these effects are more easily seen human neuroimaging: *representational churn*. When something unexpected happens, a given layer will transition from representing the predicted outcome to then representing what actually happened. This "churn" through representational states, when imaged using something like fMRI which has a long time constant of signal integration, or even faster ERP imaging along with typical aggregation and smoothing processing,

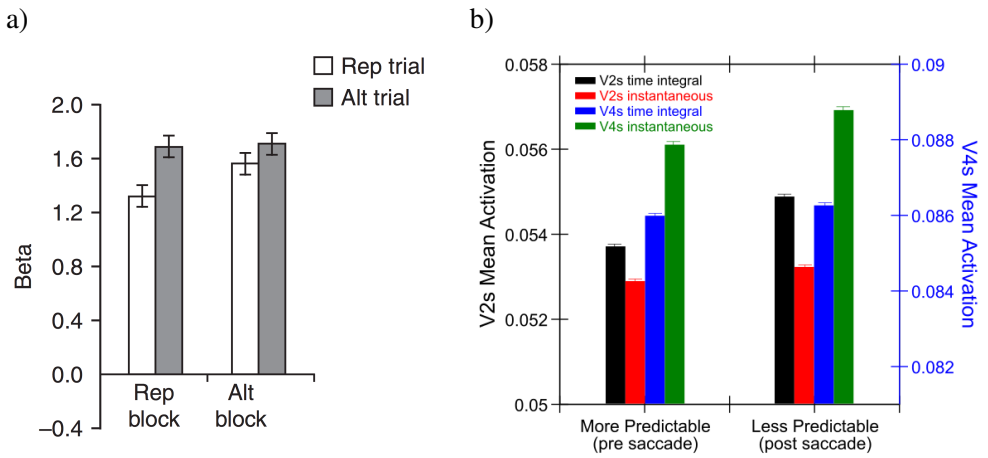

Figure 22: Activation reductions for more vs. less predictable trials a) Data from Summerfield et al. (2008) fMRI study, comparing a block-wise manipulation of probability of repetition (75% for Rep block, 25% for Alt block). Repetition suppression is enhanced when repetitions are more expected (Rep 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 because there is much less churn in V4, but it does perform its own time integral over V2.

can show up as a net overall increase in neural activation, even without instantaneous activation increasing at any given point. 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 22 shows this churn-based 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 experiences this representational churn when predictions do not quite align with the new inputs, and the time-averaged activation over this third trial is higher than when it is relatively more stable in the second trial. This is 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 churn across the two trials, and the time-averaged activations do not differ. 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 22a 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 and the churn effect, we would say that people formed stronger predictions of the face repeating in the 75% repetition block, and when it actually did repeat there was thus less churn compared to when repetitions were less frequent and predictions were weaker. 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

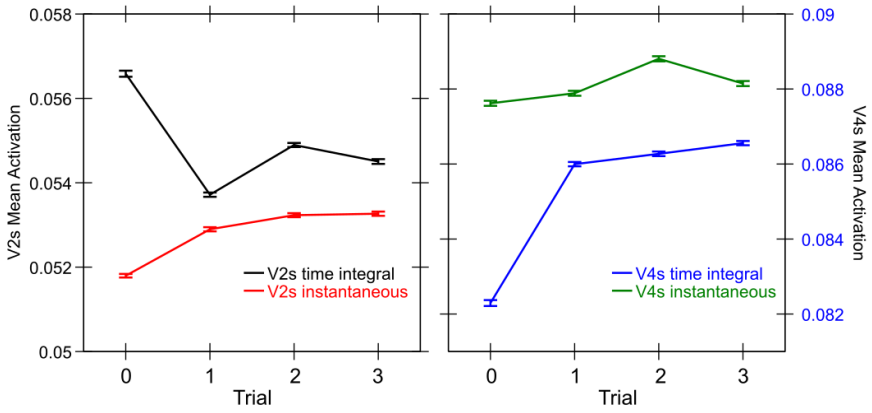

Figure 23: Time-integral and instantaneous activation across all 4 trials for V2s and V4s layers (data in previous graph comes from Trials 1,2). Overall, trial 3 should be similarly predictable as trial 1, and the activations are consistent with this. Trial 0 is highly unpredictable, and shows even higher 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.

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 representational churn.

Finally, Figure 23 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 is 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 what 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.

Attention Mechanisms in Deep / Thalamic Networks

Finally, although the focus of this paper is on predictive learning, there is another side to the DeepLeabra framework involving the ability of the very same deep / thalamic networks to modulate cortical activation, focusing attention on some elements of a scene and downregulating others. Biologically and computationally these circuits are synergistic, in that the same mechanisms serve both predictive learning and attentional functions. More generally, we think there is a larger underlying synergy, where predictions are only made about attentionally-selected objects, and, to perhaps a lesser extent, vice-versa.

Consistent with this attentional aspect of the model, there is a rapidly-growing literature on the behavioral correlates of alpha-frequency EEG power in humans, along with many demonstrations of alpha-frequency entrainment and phase effects on perception (Nunn & Osselton, 1974; Varela et al., 1981; VanRullen & Koch, 2003; Klimesch, Sauseng, & Hanslmayr, 2007; Busch, Dubois, & VanRullen, 2009; Mathewson, Fabiani, Gratton, Beck, & Lleras, 2010; Jensen & Mazaheri, 2010; VanRullen & Dubois, 2011; Palva & Palva, 2011; Rohenkohl & Nobre, 2011; Jensen et al., 2012; Jensen, Gips, Bergmann, & Bonnefond, 2014). The working hypothesis for most researchers in the field at this point is that there is a modulation of cortical inhibition at the alpha frequency, and top-down attentional mechanisms can selectively lift this inhibition, resulting in the robust finding of reduced alpha power in brain areas that are under the spotlight of attention, relative to higher alpha power in unattended areas. Most of this work has taken place in humans, and until recently the detailed biological basis for these effects have been elusive.

There is now a clear biological account emerging, based on careful laminar depth electrode recordings in monkeys, showing that alpha-frequency bursting driven by deep layer (5IB) neurons has a modulatory effect on inhibition throughout the corresponding cortical column (Dougherty et al., 2017; van Kerkoerle et al., 2014; Bortone et al., 2014; Olsen et al., 2012). Specifically, multiple researchers have found that deep-layer alpha sources of local field potential (LFP) modulate spiking of superficial-layer neurons (Dougherty et al., 2017; van Kerkoerle et al., 2014; Haegens et al., 2011; Lakatos et al., 2008; Spaak et al., 2012; Bollimunta et al., 2011; Bollimunta et al., 2008), consistent with the well-characterized effects of layer 6CT neurons (Bortone et al., 2014; Olsen et al., 2012).

Thus, the alpha cycle appears to organize both the predictive learning and attentional update dynamics, in a synergistic fashion, with the deep / thalamic network providing an outer loop to the inner-loop of superficial layer constraint-satisfaction processing. These nested loops can be thought of in terms of the widely-used expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). As elaborated below, the diffuse integrative *context* connections within the deep layer (e.g., supported by the 6CC neurons and other broad corticocortical connectivity among these deep neurons; Thomson, 2010; Thomson & Lamy, 2007), are as important for the attentional computation as they were for SRN-like temporal context as described above.

Figure 24 shows how our model captures the essential computations of the Reynolds and Heeger (2009) model in different parts of the superficial and deep layer circuits. Working backward from the 6CT modulatory layer, we posit that this layer encodes a final normalized attentional mask that has an overall multiplicative or gain-field effect on neural activations in the superficial network, which is consistent with relevant data (Bortone et al., 2014; Olsen et al., 2012; Dougherty et al., 2017; van Kerkoerle et al., 2014). Thus, where activations are strong in this layer, the corresponding superficial layer activations will remain strong, but where they are weaker, the superficial layer activations will be reduced. The normalization in 6CT occurs via inhibitory feedback circuits, both locally within layer 6 and through the TRN and TRC circuits of the thalamus (which then feed back into 6CT as well). This normalization process is affected by the 6CC layer prior to 6CT, which does the pooled integration over space and features, and then feeds into 6CT. One step prior, area 5IB combines local stimulus features and the top-down attentional inputs from higher-level areas (e.g., LIP in this case, which has been shown to support spatially-organized attentional activations; Bisley & Goldberg, 2010). Thus, all of the same essential computations that are present in the Reynolds and Heeger (2009) model are distributed across these different deep layers.

Figure 25 shows that our model captures the same key data as the Reynolds and Heeger (2009) model, where the relative balance of the enhancing vs. suppressive effects of attentional modulation can shift depending on the relative sizes of the attentional spotlight and the stimulus input (and as a function of stimulus contrast), producing the shift from contrast gain to response gain effects of attention. Thus, although there is much more work to be done here to explore the full range of attentional dynamics, this provides a solid foundation building on the well-established Reynolds and Heeger (2009) model.

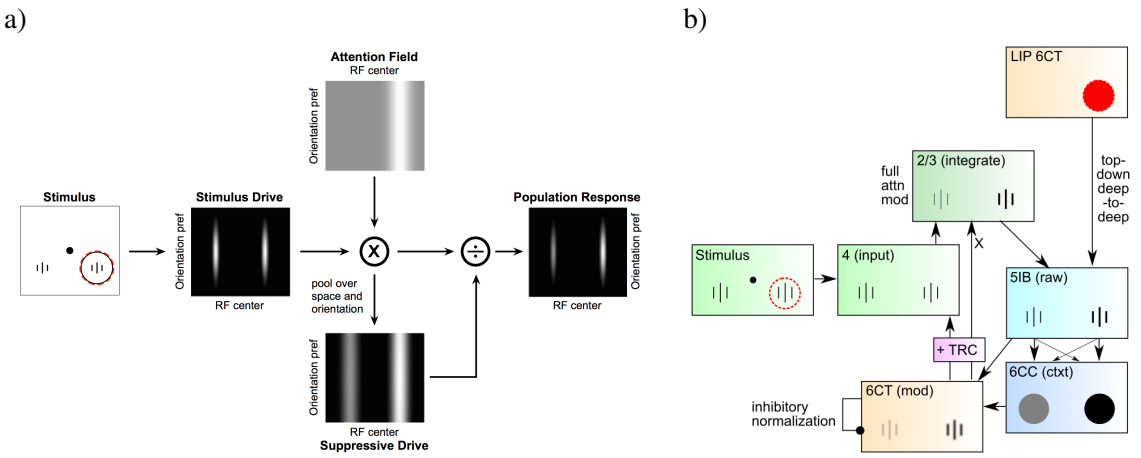

Figure 24: a) The Reynolds & Heeger (2009) computational model of pooling and normalization processes in attention. b) How attentional modulation is computed across the deep layers in DeeLeabra, in response to a top-down attentional focus (as encoded in LIP of parietal cortex). Layer 4 receives bottom-up sensory input (initially equally weighted), which then drives superficial layers (2/3), which initially do not reflect the attentional modulation (not shown). The deep 5IB neurons integrate deep-to-deep top-down attentional inputs from LIP plus the local stimulus features from 2/3, to produce the *raw* deep output, prior to the contextual normalization process. The 6CC neurons integrate across the 5IB activations (context integration or pooling). 6CT then integrates this contextual and direct activation from 5IB, to produce, for the first time in the circuit, a properly renormalized multiplicative gain-field activation pattern, with surround inhibition both within the 6CT layer and further downstream in the TRN and TRC circuit providing the critical renormalization process. These 6CT activations then modulate (multiply) the superficial-layer activations to produce *both* an increase the attended location, and a decrease for the unattended location, as shown. In the biology, this modulation affects the layer 4 inputs (not shown) as well as 2/3. Our model subsumes layer 4 into layer 2/3 neurons.

Furthermore, our model is related to the *folded-feedback* model of Grossberg (1999) (see Raizada & Grossberg, 2003 for a more elaborated version), which also posits this same kind of attentional modulation dynamic between layer 6 and the superficial layers. Interestingly, top-down attentional signals, like those coming from LIP down to lower-level visual pathways, are preferentially communicated via a network of deep-to-deep projections (Markov et al., 2014b; von Stein et al., 2000; van Kerkoerle et al., 2014).

In a future paper, we plan to apply our model to a wide range of alpha-frequency effects on perception and attention (Nunn & Osselton, 1974; Varela et al., 1981; VanRullen & Koch, 2003; Klimesch et al., 2007; Busch et al., 2009; Mathewson et al., 2010; Jensen & Mazaheri, 2010; VanRullen & Dubois, 2011; Palva & Palva, 2011; Rohenkohl & Nobre, 2011; Jensen et al., 2012; Jensen et al., 2014), to better understand how deep, thalamic, and superficial-layer dynamics interact to produce these effects, and how predictive learning and attention interact as well. Is it possible that some of these effects could be driven just by the alpha-frequency context updating in the predictive learning aspect, or are they all due to attentional modulation effects? What causes the alpha phase to reset (Calderone, Lakatos, Butler, & Castellanos, 2014), and how does the interplay between intrinsic oscillatory dynamics and external driving stimuli work? What about the effects of saccades, which also appear to reset the alpha phase (Melloni, Schwiedrzik, Rodriguez, & Singer, 2009; Paradiso, Meshi, Pisarcik, & Levine, 2012; Maldonado, Babul, Singer, Rodriguez, Berger, & Grün, 2008; Rajkai, Lakatos, Chen, Pincze, Karmos, & Schroeder, 2008; Ito, Maldonado, Singer, & Grn, 2011)?

One major goal of this work would be to provide a more satisfying integration of the inhibitory versus excitatory effects of alpha modulation (Palva & Palva, 2011, 2007; Gulbinaite, İlhan, & VanRullen, 2017). In our model (and Reynolds and Heeger (2009)), the final modulatory signal carried by layer 6CT neurons is excitatory (these are excitatory pyramidal neurons), and its multiplicative effect on other neurons is

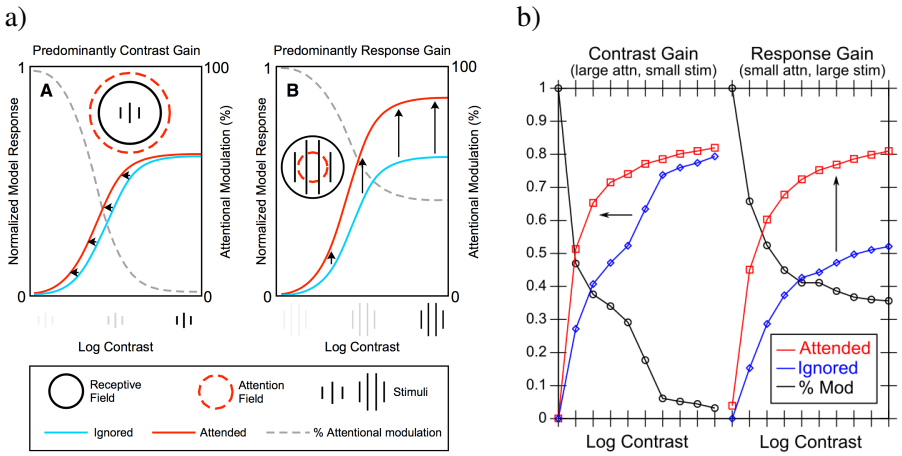

Figure 25: a) Key data accounted for by Reynolds & Heeger (2009) model, showing two qualitatively different types of attentional modulation that can emerge from the same model, as a result of differences in size of attentional spotlight relative to stimulus size. These different effects, which fit experimental data, result directly from the pooling and normalization processes, and are thus a key test of the model dynamics. b) Results from a DeepLeabra model driven by large LIP attentional top-down spotlight relative to a small bottom-up stimulus (left) versus a small LIP spotlight relative to a larger stimulus, reproducing same qualitative effects.

hypothesized to result from an interaction between excitatory and inhibitory circuits. Bortone et al. (2014) clearly demonstrate that 6CT neurons strongly activate inhibitory interneurons in layer 6 that synapse throughout the cortical column, providing a strong overall background of inhibition. However, they are also careful to emphasize that there are many excitatory synapses onto other pyramidal and thalamic TRC neurons, that can have the opposite effect, and the net overall effect is likely to depend critically on spatial topography (e.g., surround inhibition with central excitation) and also the local activity levels of the receiving neurons. Furthermore, it is not clear how the known biological mechanisms would cause the level of inhibition of superficial spiking to be a direct function of overall EEG-level alpha power, as many theories assume (e.g., Klimesch et al., 2007; Jensen et al., 2014). Instead, it is certain that gamma power increases when superficial neurons are disinhibited, which may directly remove some alpha power that they would otherwise have been contributing, and these superficial neurons may also have a consequent impact on the level of alpha synchrony in the deep layers (e.g., by affecting the timing of 5IB bursting). Thus, the causal arrow may go the other way, consistent with various issues raised by Palva and Palva (2011, 2007).

General Discussion

We have presented a comprehensive model of the visual system, driven by a core engine of predictive learning, that incorporates, and accounts for, a very wide range of data across many different levels of analysis, from low-level synaptic plasticity to systems-level organization and connectivity of the areas and pathways of the visual system (including their development), explaining how high-level invariant object representations can emerge without any explicit training signal. 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 (along with many other details enumerated throughout the paper):

- The existence of a strong synchronized, low-frequency modulation of cortex (at the alpha frequency).
- Specificity of this alpha modulation to 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.
- Early development of the *Where* (MT, LIP) 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. This is particularly true for the rather striking features of the pulvinar — although various theoretical speculations have been offered of the various properties summarized above, we are not aware of any other substantial computational models that demonstrate a clear computational, functional benefit from these properties.

Thus, overall, we argue that there is strong empirical support for our model, and no obvious gaps or major inconsistencies in the framework that would appear to invalidate it. 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 word learning and language development. 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.

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* that 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 et al., 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 central twist of this framework (e.g., Lee & Mumford, 2003; Friston, 2005; Yuille & Kersten, 2006; Friston, 2008, 2010). 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 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 the author 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 architectually 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 et al., 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 et al., 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 et al., 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 et al., 2012b; Wyatte et al., 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., 2015). 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

Many of the developmental implications of an overall predictive learning approach have been articulated by Elman et al. (1996), but a few major points bear emphasis here. 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, we computed that our sequence of *Where*-only then full model training would represent just 21 hours of real-time learning. 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 about 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, such knowledge may not be innate. Interestingly, informal conversations with Liz Spelke suggest that she would consider this kind of predictive learning account entirely compatible with her view of innate — something that is inevitable given basic environmental input combined with a carefully structured ("innate") biological system that is expectant of this experience. Certainly we agree that our model is far from a generic *tabula rasa*, and this complex interaction between learning and evolved neural structure leaves plenty of room for a nice synthesis between nativist and empiricist views.

In any case, there is a great opportunity now 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 there that a more advanced multi-modal version of our model could hopefully contribute to, especially given a recent emphasis on collecting real-world experience samples that provide considerable insight and constraints (e.g., Stevens, Gleitman, Trueswell, & Yang, 2017; Yu & Smith, 2012; Colunga & Smith, 2005; Waxman & Gelman, 2009).

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). We have provocatively referred to the pulvinar as the "silver screen of the Cartesian Theater" as a colorful way of characterizing the unique functional role of this brain area in our framework, but it also captures some of the potential broader implications. Dennett (1991) coined the term *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), and we have likewise been firm believers in a distributed, emergentist account of consciousness that attributes significant weight to the critical role of bidirectional attractor dynamics (Lamme, 2006).

However, it seems plausible that, by providing something analogous to a projection screen in the brain (updating at film-appropriate alpha frame rates no less!), the pulvinar may in fact play a critical role in organizing and coordinating diffuse brain areas around a common focus of the collaboratively-generated prediction of what will happen next. In so doing, we could say that this naturally results in 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 everything attributed to the *hard problem* (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. 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.
- 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) 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 — 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.

- 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.

- 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?

- 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, but the same prediction-generation pathway from layer 6CT to LGN does exist — perhaps it learns in a more online, continuous fashion?

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 complete, 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.

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.

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

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

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

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

Bengio, Y., 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.

Bisley, J. W., & Goldberg, M. E. (2010). Attention, intention, and priority in the parietal lobe. *Annual Review of Neuroscience*, *33*, 1–21.

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.

Busch, N. A., Dubois, J., & VanRullen, R. (2009). The phase of ongoing EEG oscillations predicts visual perception. *The Journal of Neuroscience*, *29*(24), 7869–7876.

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.

Calderone, D. J., Lakatos, P., Butler, P. D., & Castellanos, F. X. (2014). Entrainment of neural oscillations as a modifiable substrate of attention. *Trends in Cognitive Sciences*, *18*(6), 300–309.

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.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

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.

Dailey, M. N., & Cottrell, G. W. (1999). Organization of face and object recognition in modular neural network models. *Neural networks*, *12*(7-8), 1053–1074.

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.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

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.

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.

Enroth-Cugell, C., & Robson, J. G. (1966). The contrast sensitivity of retinal ganglion cells of the cat. *The Journal of physiology*, *187*(3), 517–552.

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.

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.

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.

Grossberg, S. (1999). How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex. *Spatial vision*, *12*.

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

Gulbinaite, R., İlhan, B., & VanRullen, R. (2017). The Triple-Flash Illusion Reveals a Driving Role of Alpha-Band Reverberations in Visual Perception. *Journal of Neuroscience*, *37*(30), 7219–7230.

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.

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.

Hubel, D., & Wiesel, T. N. (1962). Receptive Fields, Binocular Interaction, and Functional Architecture in the Cat's Visual Cortex. *Journal of Physiology*, *160*(1), 106–154.

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.

Ito, J., Maldonado, P., Singer, W., & Grn, S. (2011). Saccade-related modulations of neuronal excitability support synchrony of visually elicited spikes. *Cerebral Cortex*, *21*(11), 2482–2497.

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

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

Jensen, O., Gips, B., Bergmann, T. O., & Bonnefond, M. (2014). Temporal coding organized by coupled alpha and gamma oscillations prioritize visual processing. *Trends in Neurosciences*, *37*(7), 357–369.

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

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.

Kaas, J. H., & Lyon, D. C. (2007). Pulvinar contributions to the dorsal and ventral streams of visual processing in primates. *Brain Research Reviews*, *55*(2), 285–296.

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.

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

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

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

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

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., & 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.

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.

Maldonado, P., Babul, C., Singer, W., Rodriguez, E., Berger, D., & Grün, S. (2008). Synchronization of neuronal responses in primary visual cortex of monkeys viewing natural images. *Journal of Neurophysiology*, *100*(3), 1523–1532.

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., Essen, V., C, D., & 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 hierarchy: 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.

Martinez-Conde, S., Macknik, S. L., & Hubel, D. H. (2004). The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience*, *5*(3), 229–240.

Martinez-Conde, S., Otero-Millan, J., & Macknik, S. L. (2013). The impact of microsaccades on vision: Towards a unified theory of saccadic function. *Nature Reviews Neuroscience*, *14*(2), 83–96.

Masquelier, T., & Thorpe, S. J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Computational Biology*, *3*(2), 247–257.

Mathewson, K. E., Fabiani, M., Gratton, G., Beck, D. M., & Lleras, A. (2010). Rescuing stimuli from invisibility: Inducing a momentary release from visual masking with pre-target entrainment. *Cognition*, *115*(1), 186–191.

Melloni, L., Schwiedrzik, C. M., Rodriguez, E., & Singer, W. (2009). (Micro)Saccades, corollary activity and cortical oscillations. *Trends in Cognitive Sciences*, *13*(6), 239–245.

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.

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

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.

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.

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.

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. (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. (2015). 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).

O'Reilly, R. C., Wyatte, D., & Rohrlich, J. (2014c). Learning Through Time in the Thalamocortical Loops. *arXiv:1407.3432 [q-bio]*.

Palva, S., & Palva, J. M. (2007). New vistas for alpha-frequency band oscillations. *Trends in Neurosciences*, *30*(4), 150–158.

Palva, S., & Palva, J. M. (2011). Functional roles of alpha-band phase synchronization in local and large-scale cortical networks. *Frontiers in Psychology*, *2*(204), ePub only.

Paradiso, M. A., Meshi, D., Pisarcik, J., & Levine, S. (2012). Eye movements reset visual perception. *Journal of Vision*, *12*(13).

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.

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

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*.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, *46*(1), 77–105.

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.

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

Raizada, R. D. S., & Grossberg, S. (2003). Towards a theory of the laminar architecture of cerebral cortex: Computational clues from the visual system. *Cerebral cortex*, *13*(1).

Rajkai, C., Lakatos, P., Chen, C.-M., Pincze, Z., Karmos, G., & Schroeder, C. E. (2008). Transient cortical excitation at the onset of visual fixation. *Cerebral Cortex*, *18*(1), 200–209.

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.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*(2), 168–185.

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

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

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

Rockland, K. S. (1998a). Complex microstructures of sensory cortical connections. *Current opinion in neurobiology*, *8*, 545.

Rockland, K. S. (1998b). 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.

Rohenkohl, G., & Nobre, A. C. (2011). Alpha oscillations related to anticipatory attention follow temporal expectations. *The Journal of Neuroscience*, *31*(40), 14076–14084.

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.

Rumelhart, D. E., & McClelland, J. L. (1986). PDP Models and General Issues in Cognitive Science. In D. E. Rumelhart, J. L. McClelland, & P. R. Group (Eds.), *Parallel Distributed Processing. Volume 1: Foundations* (pp. 110–146). Cambridge, MA: MIT Press.

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.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, *45*(11), 2673–2681.

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).

Shrager, J., & Johnson, M. H. (1996). Dynamic Plasticity Influences the Emergence of Function in a Simple Cortical Array. *Neural Networks*, *9*(7), 1119–1129.

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

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

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.

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.

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., & Dubois, J. (2011). The psychophysics of brain rhythms. *Frontiers in Psychology*, *2*(203).

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. Google-Books-ID: cSjEAgAAQBAJ.

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.

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, *51*(2), 167–194.

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.

Young, R. A. (1987). The Gaussian derivative model for spatial vision: I. Retinal mechanisms. *Spatial Vision*, *2*(4), 273–293.

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.

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.

Appendix: Computational Model Details

TODO: needs updating

This appendix provides more information about the object recognition 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`.

Structure of the model

The detailed flow of activation according to these principles is illustrated in Figure 26, showing stimulus information coming into area V1, proceeding up to area V2 via the superficial network, and back down to area V1, again in the superficial network. All of this is modulated by ongoing layer 6CT activations driven from prior network state. Toward the end of the first 100 msec alpha cycle, the 5IB bursting from area V1 then drives a plus phase activation state in the TRC neurons of V2, which then propagates up through the V2 superficial and deep networks. Meanwhile, the V2 5IB neurons fire and drive a wave of new activation through the rest of the V2 deep network (6CC to 6CT), corresponding to an update in the attentional and temporal context state information for the V2 layer. The same update also happens in V1. During the next alpha cycle, these updated deep network activation states are then continuously communicated by the 6CT neurons to the TRC neurons in the thalamus (with TRN contrast enhancement), resulting in a minus phase expectation for the subsequent plus-phase state over the TRC's, and also a direct broadcast of the 6CT activation state up to the superficial cortex which serves to multiplicatively modulate the activation states there, producing an attentional modulation effect.

This information is largely the same as for the model we are building upon (O'Reilly et al., 2013), but we have changed the scale of the model and several other features to incorporate the DeepLeabra learning mechanisms.

Early Visual Image Processing

In the mammalian brain, the retina and LGN compress the visual input into an efficient contrast-coded representation using center-surround contrast filters that are radially symmetric (which can be nicely approximated by the difference of two Gaussians, Enroth-Cugell & Robson, 1966; Young, 1987). Then area V1 encodes orientation and other features building upon this basic contrast-enhanced input (Hubel & Wiesel, 1962). We compress this chain of filters into a single step by using oriented Gabor filters, which are defined as a Gaussian-shaped spatial weighting multiplying a planar sine wave oriented in a given direction:

$$g(x, y) = e^{-\left(\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2}\right)} \sin\left(\frac{2\pi y}{\lambda}\right) \tag{2}$$

where the sine wave moves along the y axis (corresponding to a horizontal orientation tuning), and the Gaussian has differential width terms (σ_x, σ_y) for each axis. To obtain different orientations, the coordinates x,y are rotated by a given angle θ relative to the original coordinates of the filter (x', y'):

$$\begin{aligned} x &= x' \cos(\theta) - y' \sin(\theta) \\ y &= y' \cos(\theta) + x' \sin(\theta) \end{aligned} \tag{3}$$

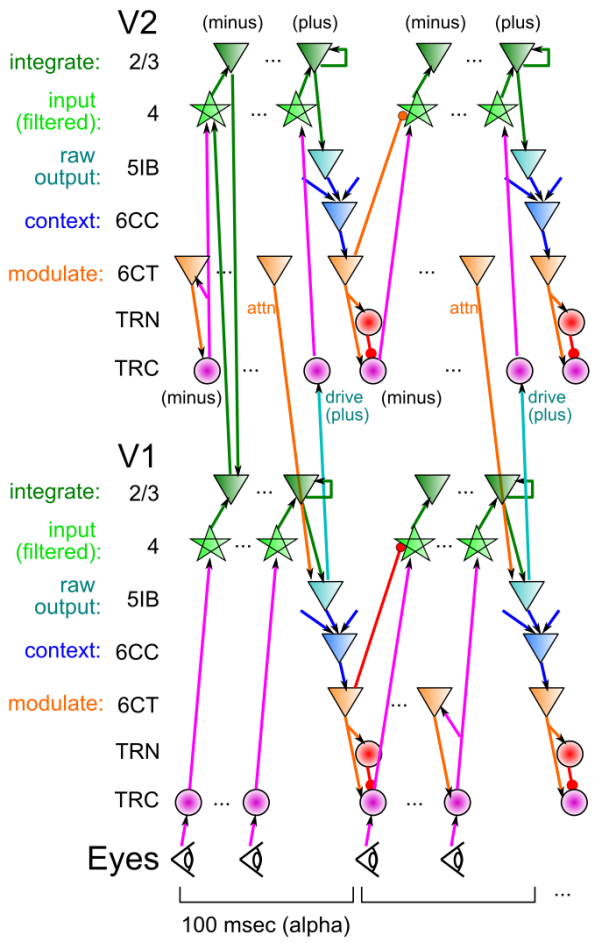

Figure 26: The temporal evolution of information flow in a DeepLeabra model across V1 and V2 layers. Information flows in from V1 to V2 and back in the superficial layer network (in green), while the deep network drives attentional modulation and temporal context information for predictive auto-encoder learning. Computationally, the superficial neurons serve to integrate information through constraint satisfaction, and they receive attentionally-filtered signals from layer 4 neurons that are modulated via 6CT (corticothalamic) projections. The 5IB neurons at the start of the deep layer network represent the most salient, raw (prior to contextual normalization) output of a given area, and the 6CC neurons integrate across this raw output to produce a context representation that supports the ability to make predictions about the next input state, and also drives renormalization of the 6CT attentional modulation signal.

The filter is always normalized to a zero sum in the discrete kernel that is actually used, to ensure that a uniform illumination produces no activation.

In line with other established models of object recognition in cortex (e.g., Wallis & Rolls, 1997; Riesenhuber & Poggio, 1999; Dailey & Cottrell, 1999; Masquelier & Thorpe, 2007), these filtering operations provide a reasonable approximation to the coding properties of V1 simple cells. The model processes each image at two different spatial frequencies (SF), "high" and "medium", each of which employs 4 orientations of tuning, times 2 for on vs. off-center polarity. The high-SF pathway uses Gabor filters rendered on a 6x6 pixel kernel, with a wavelength $\lambda = 6$, and Gaussian width terms of $\sigma_x = 1.8$ and $\sigma_y = 1.2$. The medium-SF pathway uses Gabor filters that are twice as large (12x12 kernel, wavelength $\lambda = 12$, and Gaussian width terms of $\sigma_x = 3.6$ and $\sigma_y = 2.4$). These filters are applied in a half-overlapping fashion to the image, such that adjacent V1 simple units process spatial locations that are one half wavelength ($\frac{\lambda}{2}$) away from their neighbors.

The input "retina" resolution is only 24x24 pixels, and the high frequency V1 simple filters are computed centered on each pixel (spacing = 1), producing a 24x24x8 (where 8 refers to 4 orientations x 2 polarities) dimensional output, while the medium frequency have a spacing of 2 and produce a 12x12x8 dimensional output.

The output of the V1 simple cells is computed using the FFFB inhibitory dynamics and point-neuron activation function of Leabra (described below), applied to a net input that is the positive-rectified (values less than 0 are clipped to 0) result of convolving the Gabor kernel with the input image. There are two levels of inhibitory competition — the primary is within the group of different orientation and polarity tunings for the same spatial location (i.e., 8 units = 4 orientations x 2 polarities). This *unit group* level competition may reflect competition at the level of the *hypercolumn* in the brain. This inhibition is computed with a gi value of 2.0. The secondary level of competition involves a spread of the unit-group level competition across the entire layer of such units, with a discounted gain multiplier and a MAX operation such that the stronger of the unit-group or discounted layer-level competition holds (see FFFB section of Leabra algorithm section for details).

Structure of Higher Layers (Extrastriate, Inferotemporal, Output, Semantics)

Proceeding from the V1 simple inputs at the two different spatial frequencies (high and medium), the model captures the general response properties of extrastriate cortex (V2, V3, V4) and inferotemporal (IT) cortex. As a purely computational convenience in configuring the network, the model's V2 layers remain spatial-frequency specific (in the brain, we would expect these to all be intermixed), which then merge into unitary V3, V4, and IT layers, and IT then feeds into a naming output layer, and a semantics output layer (Figure **??**). All connections are bidirectional, except those from V2 back to V1 in the implicit prediction case — for explicit prediction, V2 does project back to V1 to generate the minus phase expectation.

All of the units are allowed to learn based on the Leabra learning mechanism (described in the next section). Once trained, the single model can discriminate all trained object categories — in contrast, other prevalent feedforward models (e.g., Riesenhuber & Poggio, 1999; Masquelier & Thorpe, 2007) use binary classifiers that would require *N* classifiers to differentiate among *N* categories. Thus, the overall solution to the invariant object recognition problem that the model develops is qualitatively similar with the prevalent feedforward models, yet is also realizable using a homogeneous, biologically plausible set of mechanisms.

Here are the detailed parameters for each layer in the network (note that 15-25% activity levels is the default for Leabra models of the cortex, based on biological estimates; O'Reilly & Munakata, 2000; O'Reilly et al., 2012):

- **V2**: 25 units per unit group / hypercolumn, arranged into a 12x12 topographical grid for the high spatial frequency layer, and a 6x6 grid for the medium spatial frequency layer. Each unit receives from a topographically-corresponding 4x4 grid of V1 unit groups, with 1/2 overlap among neighboring unit groups. FFFB inhibition produces roughly 10% activity within each unit group.
- **V3:** 576 total units receiving a full projection from all V2 units (and projecting bidirectionally back to them), and from the V2_sum layer that summarizes the V2 unit groups with a single unit.
- **V4:** 64 units per unit group / hypercolumn, arranged into a 3x3 topographical grid, and receiving a retinotopically-organized projection from 4x4 V2h unit groups, and from 8x8 V2m unit groups, half overlapping as before.
- **IT**: 200 total units receiving a full projection from all of the V3 and V4 units (and projecting bidirectionally back to them), with a 15% kWTA activity level (no unit group sub-structure)

- **Naming Output**: 200 units receiving a full projection from the IT (and projecting completely back to it), with a kWTA activity level of 1%. This localist (single active unit) representation of output names is a computational simplification, standing in for the full phonological production pathways.

Model Algorithms

The model was implemented using the Leabra framework, which is described in detail in O'Reilly et al. (2012), O'Reilly and Munakata (2000), O'Reilly (2001), and summarized here. See Table 2 for a listing of parameter values, nearly all of which are at their default settings. These same parameters and equations 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.

This version of Leabra contains three primary differences from the original (O'Reilly & Munakata, 2000): the activation function is slightly different, in a way that allows units to more accurately reflect their graded excitatory input drive, the inhibition function is much simpler and more biologically realistic, and the learning rule takes a more continuous form involving contrasts between values integrated over different time frames (i.e., with different time constants), which also produces a combination of error-driven and self-organizing learning within the same simple mathematical framework. These modifications are described in detail in an updated version of the O'Reilly and Munakata (2000) textbook, in O'Reilly et al. (2012). This new learning algorithm goes by the acronym of XCAL (temporally eXtended Contrastive Attractor Learning), and it replaces the combination of Contrastive Hebbian Learning (CHL) and standard Hebbian learning used in the original Leabra framework.

Pseudocode

The pseudocode for Leabra is given here, showing exactly how the pieces of the algorithm described in more detail in the subsequent sections fit together. The individual steps are repeated for each event (trial), which can be broken down into a *minus* and *plus* phase, followed by a synaptic weight updating function. Generally speaking, the minus phase represents the system's expectation for a given input and the plus phase represents the observation of the outcome. The difference between these two phases is then used to compute the updating function that drives learning. Furthermore, each phase contains a period of *settling* (measured in *cycles*) during which the activation values of each unit are updated taking into account the previous state of the network. Some units are *clamped*, or have fixed activation values and are not subject to this updating rule (e.g., V1 input in the minus phase, V1 input and Output in the plus phase).

Outer loop: For each event (trial) in an epoch:

1. Iterate over minus and plus phases of settling for each event.
   (a) At start of settling, for all units:
      i. Initialize all state variables (activation, V_m, etc).
      ii. Clamp external patterns (V1 input in minus phase, V1 input & Output in plus phase).
   (b) During each cycle of settling, for all non-clamped units:
      i. Compute excitatory netinput ($g_e(t)$ or η_j, eq 6).
      ii. Compute FFFB inhibition for each layer, based on average net input and activation levels within the layer (eq 12)
      iii. Compute point-neuron activation combining excitatory input and inhibition (eq 4).
      iv. Update time-averaged activation values (short, medium, long) for use in learning.

Parameter	Value	Parameter	Value
E_l	0.30	$\overline{g_l}$	0.10
E_i	0.25	$\overline{g_i}$	1.00
E_e	1.00	$\overline{g_e}$	1.00
V_{rest}	0.30	Θ	0.50
τ	.3	γ	80

Table 2: Parameters for the simulation (see equations in text for explanations of parameters). All are standard default parameter values.

2. After both phases update the weights, for all connections:

   (a) Compute XCAL learning as function of short, medium, and long time averages.

   (b) Increment the weights according to net weight change.

Point Neuron Activation Function

Leabra uses a *point neuron* activation function that models the electrophysiological properties of real neurons, while simplifying their geometry to a single point. This function is nearly as simple computationally as the standard sigmoidal activation function, but the more biologically-based implementation makes it considerably easier to model inhibitory competition, as described below. Further, using this function enables cognitive models to be more easily related to more physiologically detailed simulations, thereby facilitating bridge-building between biology and cognition. We use normalized units where the unit of time is 1 msec, the unit of electrical potential is 0.1 V (with an offset of -0.1 for membrane potentials and related terms, such that their normal range stays within the $[0, 1]$ normalized bounds), and the unit of current is $1.0x10^{-8}$.

The membrane potential V_m is updated as a function of ionic conductances g with reversal (driving) potentials E as follows:

$$\Delta V_m(t) = \tau \sum_c g_c(t)\overline{g_c}(E_c - V_m(t)) \tag{4}$$

with 3 channels (c) corresponding to: e excitatory input; l leak current; and i inhibitory input. Following electrophysiological convention, the overall conductance is decomposed into a time-varying component $g_c(t)$ computed as a function of the dynamic state of the network, and a constant $\overline{g_c}$ that controls the relative influence of the different conductances. The equilibrium potential can be written in a simplified form by setting the excitatory driving potential (E_e) to 1 and the leak and inhibitory driving potentials (E_l and E_i) of 0:

$$V_m^\infty = \frac{g_e\overline{g_e}}{g_e\overline{g_e} + g_l\overline{g_l} + g_i\overline{g_i}} \tag{5}$$

which shows that the neuron is computing a balance between excitation and the opposing forces of leak and inhibition. This equilibrium form of the equation can be understood in terms of a Bayesian decision making framework (O'Reilly & Munakata, 2000).

The excitatory net input/conductance $g_e(t)$ or η_j is computed as the proportion of open excitatory channels as a function of sending activations times the weight values:

$$\eta_j = g_e(t) = \langle x_i w_{ij} \rangle = \frac{1}{n} \sum_i x_i w_{ij} \tag{6}$$

The inhibitory conductance is computed via the kWTA function described in the next section, and leak is a constant.

In its discrete spiking mode, Leabra implements exactly the AdEx (adaptive exponential) model (Brette & Gerstner, 2005), which has been found through various competitions to provide an excellent fit to the actual firing properties of cortical pyramidal neurons (Gerstner & Naud, 2009), while remaining simple and efficient to implement. However, we typically use a rate-code approximation to discrete firing, which produces smoother more deterministic activation dynamics, while capturing the overall firing rate behavior of the discrete spiking model.

We recently discovered that our previous strategy of computing a rate-code graded activation value directly from the membrane potential is problematic, because the mapping between V_m and mean firing rate is not a one-to-one function in the AdEx model. Instead, we have found that a very accurate approximation to the discrete spiking rate can be obtained by comparing the excitatory net input directly with the effective computed amount of net input required to get the neuron firing over threshold (g_e^Θ), where the threshold is indicated by Θ:

$$g_e^\Theta = \frac{g_i \bar{g}_i (E_i - V_m^\Theta) + \bar{g}_l (E_l - V_m^\Theta)}{\bar{g}_e (V_m^\Theta - E_e)} \tag{7}$$

$$y_j(t) \propto g_e(t) - g_e^\Theta \tag{8}$$

where $y_j(t)$ is the firing rate output of the unit.

We continue to use the Noisy X-over-X-plus-1 (NXX1) function, which starts out with a nearly linear function, followed by a saturating nonlinearity:

$$y_j(t) = \frac{1}{\left(1 + \frac{1}{\gamma [g_e(t) - g_e^\Theta]_+}\right)} \tag{9}$$

where γ is a gain parameter, and $[x]_+$ is a threshold function that returns 0 if $x < 0$ and x if $x > 0$. Note that if it returns 0, we assume $y_j(t) = 0$, to avoid dividing by 0. As it is, this function has a very sharp threshold, which interferes with graded learning learning mechanisms (e.g., gradient descent). To produce a less discontinuous deterministic function with a softer threshold, the function is convolved with a Gaussian noise kernel ($\mu = 0$, $\sigma = .005$), which reflects the intrinsic processing noise of biological neurons:

$$y_j^*(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2/(2\sigma^2)} y_j(z - x) dz \tag{10}$$

where x represents the $[g_e(t) - g_e^\Theta]_+$ value, and $y_j^*(x)$ is the noise-convolved activation for that value. In the simulation, this function is implemented using a numerical lookup table.

There is just one last problem with the equations as written above: They don't evolve over time in a graded fashion. In contrast, the Vm value does evolve in a graded fashion by virtue of being iteratively computed, where it incrementally approaches the equilibrium value over a number of time steps of updating. Instead the activation produced by the above equations goes directly to its equilibrium value very quickly, because it is calculated based on excitatory conductance and does not take into account the sluggishness with which changes in conductance lead to changes in membrane potentials (due to capacitance).

To introduce graded iterative dynamics into the activation function, we just use the activation value ($y^*(x)$) from the above equation as a "driving force" to an iterative temporally-extended update equation:

$$y_j(t) = y_j(t-1) + dt_{vm} \left(y_j^*(t) - y_j(t-1) \right) \tag{11}$$

This causes the actual final rate code activation output at the current time t, $y(t)$ to iteratively approach the driving value given by $y^*(x)$, with the same time constant dt_{vm} that is used in updating the membrane potential. In practice this works extremely well, better than any prior activation function used with Leabra.

FFFB Inhibition

Leabra computes a layer-level inhibition conductance value based on a combination of feed-forward (FF) and feed-back (FB) dynamics. This is an advance over the more explicit kWTA (k-Winners-Take-All) function that was used previously, though it achieves roughly the same overall kWTA behavior, with a much simpler, more efficient, and biologically plausible formulation. The FF component is based directly on the average excitatory net input coming into the layer ($< \eta >$), and the FB component is based on the average activation of units within the layer ($< act >$). Remarkably, fixed gain factors on each of these terms, together with simple time integration of the FB term to prevent oscillations, produces results that are overall comparable to the kWTA dynamics, except that the activations of units in the layer retain more of a proportional response to their overall level of excitatory drive, which is desirable in many cases.

FFFB is conceptually just the sum of the FF and FB components, each with their own ff and fb gain factors, with an overall gain factor (gi) applied to both:

$$g_i = \text{gi}\,(\text{ff}[< \eta > - \text{ff0}]_+ + \text{fb} < act >) \tag{12}$$

where $[x]_+$ indicates the positive part of whatever it contains — anything negative truncates to zero. It is important to have a small offset on the FF component, parameterized by ff0 which is typically .1 — this delays the onset of inhibition and allows the neurons to get a little bit active first. To minimize oscillations, the feedback component needs to be time integrated, with a fast time constant of .7 — just a simple exponential approach to the driving fb inhibition value was used:

$$fb_i(t) = fb_i(t-1) + dt\,(\text{fb} < act > - fb_i(t-1)) \tag{13}$$

Typically ff is set to 1.0, fb is 0.5, and the overall gain (gi) is manipulated to achieve desired activity levels — typically it is around 2.2 or so.

XCAL Learning

The full treatment of the new XCAL version of learning in Leabra is presented in O'Reilly et al. (2012), but the basic equations and a brief motivation for them are presented here.

In the original Leabra algorithm, learning was the sum of two terms: an error-driven component and a Hebbian self-organizing component. In the new XCAL formulation, the error-driven and self-organizing factors emerge out of a single learning rule, which was derived from a biologically detailed model of synaptic plasticity by Urakubo et al. (Urakubo et al., 2008), and is closely related to the Bienenstock, Cooper & Munro (BCM) algorithm (Bienenstock et al., 1982). In BCM, a Hebbian-like sender-receiver activation product term is modulated by the extent to which the receiving unit is above or below a long-term running average activation value:

$$\Delta_{bcm} w_{ij} = xy(y - \langle y^2 \rangle) \tag{14}$$

(x = sender activation, y = receiver activation, and $\langle y^2 \rangle$ = long-term average of squared receiver activation). The long-term average value acts like a dynamic plasticity threshold, and causes less-active units to increase their weights, while more-active units tend to decrease theirs (i.e., a classic homeostatic function). This form of learning resembles Hebbian learning in several respects, but can learn higher-order statistics, whereas Hebbian learning is more constrained to extract low-order correlational statistics. Furthermore, the BCM model may provide a better account of various experimental data, such as monocular deprivation experiments (Cooper et al., 2004).

TODO: add the urakubo model mechanisms fig here!

The Leabra XCAL learning rule is based on a contrast between a sender-receiver activation product term (shown initially as just xy — relevant time scales of averaging for this term are elaborated below) and a

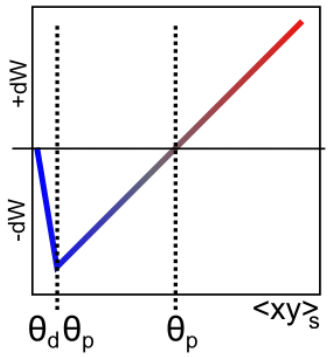

Figure 27: XCAL dWt function, shown with $\theta_p = 0.5$, which determines the cross-over point between negative and positive weight changes, and $\theta_p\theta_d$ determines the inflection point at the left where the curve goes from a negative slope to a positive slope. This function fits the results of the highly detailed Urakubo et al (Urakubo et al., 2008) model, with a correlation value of $r = 0.89$.

dynamic plasticity threshold θ_p (also elaborated below), which are integrated in the XCAL learning function (Figure 27):

$$\Delta_{xcal} w_{ij} = f_{xcal}(xy, \theta_p) \tag{15}$$

where the XCAL learning function was derived by fitting a piecewise-linear function to the Urakubo et al (Urakubo et al., 2008) simulation results based on synaptic drive levels (sender and receiver firing rates; the resulting fit was very good, with a correlation of $r = 0.89$):

$$f_{xcal}(xy, \theta_p) = \begin{cases} (xy - \theta_p) & \text{if } xy > \theta_p\theta_d \\ -xy(1 - \theta_d)/\theta_d & \text{otherwise} \end{cases} \tag{16}$$

($\theta_d = .1$ is a constant that determines the point where the function reverses back toward zero within the weight decrease regime — this reversalpoint occurs at $\theta_p\theta_d$, so that it adapts according to the dynamic θ_p value).

The BCM equation produces a curved quadratic function that has the same qualitative shape as the XCAL function (Figure 27). A critical feature of these functions is that they go to 0 as the synaptic activity goes to 0, which is in accord with available data, and that they exhibit a crossover point from LTD to LTP as a function of synaptic drive (which is represented biologically by intracellular Calcium levels). A nice advantage of the linear XCAL function is that, to first approximation, it is just computing the subtraction $xy - \theta_p$.

To achieve full error-driven learning within this XCAL framework, we just need to ensure that the core subtraction represents an error-driven learning term. In the original Leabra, error-driven learning via the Contrastive Hebbian Learning algorithm (CHL) was computed as:

$$\Delta_{chl} = x^+ y^+ - x^- y^- \tag{17}$$

where the superscripts represent the plus $(+)$ and minus $(-)$ phases. This equation was shown to compute the same error gradient as the backpropagation algorithm, subject to symmetry and a 2nd-order numerical integration technique known as the midpoint method, based the generalized recirculation algorithm (GeneRec; (O'Reilly, 1996)). In XCAL, we replace these values with time-averaged activations computed over different time scales:

- **s** = short time scale, reflecting the most recent state of neural activity (e.g., past 100-200 msec). This is considered the "plus phase" — it represents the *outcome* information on the current trial, and in general should be more correct than the medium time scale.

- **m =** medium time scale, which integrates over an entire psychological "trial" of roughly a second or so — this value contains a mixture of the "minus phase" and the "plus phase", but in contrasting it with the short value, it plays the role of the minus phase value, or expectation about what the system thought should have happened on the current trial.
- **l =** long time scale, which integrates over hours to days of processing — this is the BCM-like threshold term.

Thus, the error-driven aspect of XCAL learning is driven essentially by the following term:

$$\Delta_{xcal-err} w_{ij} = f_{xcal}(x_s y_s, x_m y_m) \tag{18}$$

However, consider the case where either of the short term values (x_s or y_s) is 0, while both of the medium-term values are > 0 — from an error-driven learning perspective, this should result in a significant weight decrease, but because the XCAL function goes back to 0 when the input drive term is 0, the result is no weight change at all. To remedy this situation, we assume that the short-term value actually retains a small trace of the medium-term value:

$$\Delta_{xcal-err} w_{ij} = f_{xcal}(\kappa x_s y_s + (1-\kappa) x_m y_m, x_m y_m) \tag{19}$$

(where $\kappa = .9$, such that only .1 of the medium-term averages are incorporated into the effective short-term average).

The self-organizing aspect of XCAL is driven by comparing this same synaptic drive term to a longer-term average, as in the BCM algorithm:

$$\Delta_{xcal-so} w_{ij} = f_{xcal}(\kappa x_s y_s + (1-\kappa) x_m y_m, \gamma_l y_l) \tag{20}$$

where $\gamma_l = 3$ is a constant that scales the long-term average threshold term (due to sparse activation levels, these long-term averages tend to be rather low, so the larger gain multiplier is necessary to make this term relevant whenever the units actually are active and adapting their weights).

Combining both of these forms of learning in the full XCAL learning rule amounts to computing an aggregate θ_p threshold that reflects a combination of both the self-organizing long-term average, and the medium-term minus-phase like average:

$$\Delta_{xcal} w_{ij} = f_{xcal}(\kappa x_s y_s + (1-\kappa) x_m y_m, \lambda \gamma y_l + (1-\lambda) x_m y_m) \tag{21}$$

where $\lambda = .01$ is a weighting factor determining the mixture of self-organizing and error-driven learning influences (as was the case with standard Leabra, the balance of error-driven and self-organizing is heavily weighted toward error driven, because error-gradients are often quite weak in comparison with local statistical information that the self-organizing system encodes).

The weight changes are subject to a soft-weight bounding to keep within the $0-1$ range:

$$\Delta_{sb} w_{ij} = [\Delta_{xcal}]_+ (1 - w_{ij}) + [\Delta_{xcal}]_- w_{ij} \tag{22}$$

where the $[]_+$ and $[]_-$ operators extract positive values or negative-values (respectively), otherwise 0.

Finally, as in the original Leabra model, the weights are subject to contrast enhancement, which magnifies the stronger weights and shrinks the smaller ones in a parametric, continuous fashion. This contrast enhancement is achieved by passing the linear weight values computed by the learning rule through a sigmoidal nonlinearity of the following form:

$$\hat{w}_{ij} = \frac{1}{1 + \left(\frac{w_{ij}}{\theta(1-w_{ij})}\right)^{-\gamma}} \tag{23}$$

where $\hat{w}_{ij}$ is the contrast-enhanced weight value, and the sigmoidal function is parameterized by an offset θ and a gain γ (standard defaults of 1 and 6, respectively, used here).

TI Context

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

$$\eta_{ti} = \langle x_i w_{ij} \rangle = \frac{1}{n} \sum_i x_i w_{ij} \tag{24}$$

This net input is then added in with the standard net input (equation 6) at each cycle of processing.

Learning of the context weights occurs through the superficial neuron's error signal, as discussed in the main text, with the sending activation being the *prior* time step's plus phase activation. We use a simple delta rule given that the context representations are static throughout the trial.

$$\Delta_{ti} w_{ij} = x_{t-1} \left(y_j^+ - y_j^- \right) \tag{25}$$

where y is the superficial neural activation in the plus and minus phases, as denoted, and x_{t-1} is the sending activation from the prior plus phase.

In general, these context projections exist for all standard projection pathways in the model, in addition to the self-context of the layer onto itself.