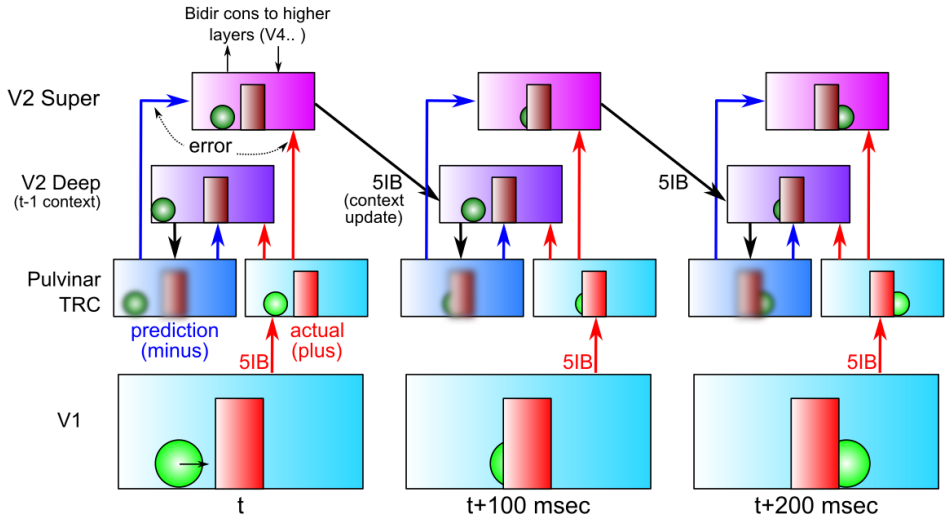
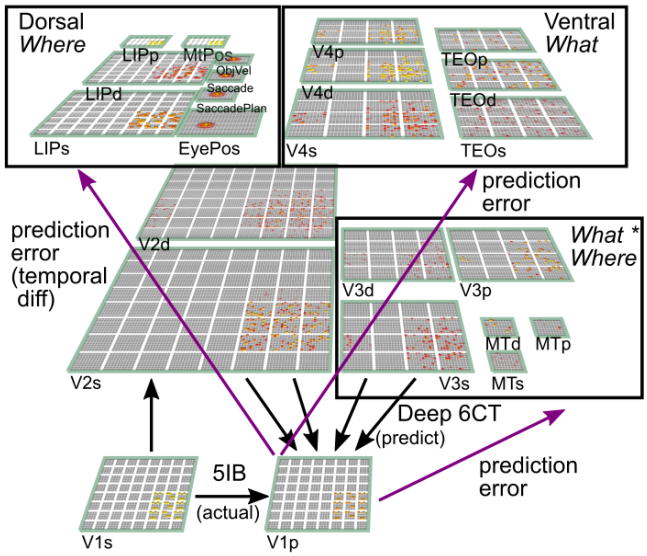
Deep predictive learning in vision

Randall C. O'Reilly, Dean R. Wyatte, and John Rohrlich
Department of Psychology and Neuroscience, University of Colorado Boulder

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



Schematic illustration of the temporal evolution of predicting visual sequences, over a period of three alpha cycles of 100 msec each. During each cycle, the V2 Deep layer uses the prior 100 msec of context information to generate a prediction or expectation (minus phase) over the pulvinar thalamic relay cell (TRC) units of what will come in next via the 5IB strong driver inputs from V1, which herald the next plus or target phase of learning. Error-driven learning occurs as a function of the temporal difference between the plus and minus activation states, in both superficial and deep networks, via the TRC projections into these networks.



The 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). *s* suffix = superficial layer, *d* = deep layer.

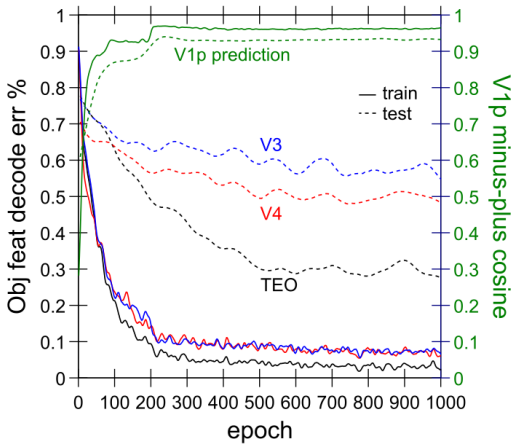


Figure 1: Learning curves for full model, showing accuracy (proportion error) in decoding the object features from each of 3 different layers (V3, V4, TEO), and overall prediction accuracy in terms of minus vs. plus phase cosine over the V1p pulvinar layer, at trial 3 (the last trial), which is nearly 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.