Deep Predictive Learning as a Model of Human Learning

Randall C. O'Reilly[1], Jacob L. Russin[1], & John Rohrlich[1]

[1]*Departments of Psychology and Computer Science, Center for Neuroscience, University of California, Davis*

Understanding the principles underlying the power of human learning is a widely held goal of research in machine learning, yet these models unrealistically rely on massive human-labeled datasets. We present a biologically based model of predictive learning, which generates predictions at the alpha frequency (100 msec), and learns from prediction errors, requiring no labeled inputs. Distinctive patterns of connectivity between the neocortex and thalamus drive alternating top-down prediction and bottom-up outcome representations over the pulvinar nucleus, with the temporal difference driving error-driven learning throughout neocortex. However, can it learn abstractions that go beyond the surface input level? Our model does, in ways that match both human category representations and monkey electrophysiology, while comparison models lacking biological features do not.

The success of deep convolutional neural networks (DCNN's) [1–3] in object recognition and many other domains raises the question of how well they model human learning, at both neural and cognitive levels. Although the engine of these models, error backpropagation [4], has long been questioned on biological grounds [5], various related biologically plausible mechanisms have been proposed [6–8]. However, the need for massive amounts of labeled data still makes these models cognitively implausible: non-human primates and infants learn to recognize and categorize objects without the benefit of such labeled data [9]. An alternative, biologically plausible approach is to use *predictive error-driven learning*, where error signals arise from differences between a prediction

of what will happen next, and what actually does occur [10,11]. In principle, all that this requires is: 1) for events to unfold over time; 2) a learning system that is somehow organized to generate predictions of these events; and 3) a biological mechanism that learns from prediction errors. Furthermore, a system that learned to accurately predict complex real-world events would require considerable knowledge to have been acquired in the process of so doing, and thus there is reason to believe that predictive learning could power sophisticated, important forms of developmental learning.

Here we show that canonical circuits between the neocortex and thalamus have several distinctive properties that directly support predictive error-driven learning. When implemented in a computational model employing a biologically plausible form of error backpropagation [6,12,13], along with several other important properties of the mammalian visual system, the model learns to systematically categorize 3D objects according to invariant shape properties. Furthermore, this category structure matches human judgments of these same objects, and is consistent with neural representations in inferotemporal (IT) cortex in primates. Comparison models with the same architecture but using standard non-biological error-backpropagation learning, and models using the state-of-the-art *PredNet* predictive learning architecture [14], support the idea that predictive learning is useful for shaping internal representations, but these models do not learn much beyond the similarities present at the lowest visual levels. Thus, we argue that incorporating biological properties of the brain can potentially provide a better understanding of human learning at multiple levels relative to existing DCCN models.

Motivated by biological evidence, we hypothesize that sensory predictions in posterior neocortex are generated roughly every 100 msec (i.e., the *alpha* rhythm, 10 Hz), by neurons in the

deep layers of the neocortex that project to the pulvinar nucleus of the thalamus (Figure a) [15]. The pulvinar represents this top-down prediction for roughly 75 msec of the alpha cycle as it develops, after which point the layer 5IB intrinsic-bursting neurons send strong, bottom-up driving input to the pulvinar, representing the actual sensory stimulus [16]. These 5IB neurons burst at the alpha frequency, determining the overall timing of the predictive learning cycle, along with other dynamic parameters of the thalamocortical circuit [17–19]. The prediction error is implicit in the temporal difference between these two periods of activity within the alpha cycle over the pulvinar, which is consistent with the biologically plausible form of error-driven cortical learning used in our models [6]. The pulvinar sends broad projections back up to all of the areas that drive top-down predictions into it [20,21], thus broadcasting this error signal to drive local synaptic plasticity in the neocortex. This mathematically approximates gradient descent to minimize overall prediction errors. This computational framework makes sense of otherwise puzzling anatomical and physiological properties of the cortical and thalamic networks [16], and is consistent with a wide range of detailed neural and behavioral data regarding the effects of the alpha rhythm on learning and perception [22–25]. It has many testable differences from other existing theories of predictive learning that have been proposed over the years, at varying levels of biological detail [26–29].

A critical question for predictive learning is whether it can develop high-level, abstract ways of representing the raw sensory inputs, while learning from nothing but predicting these low-level visual inputs. For example, can predictive learning really eliminate the need for human-labeled image datasets where abstract category information is explicitly used to train object recognition models via error-backpropagation? From a cognitive perspective, there is considerable evidence that non-verbal primates, and pre-verbal human infants, naturally develop abstract categorical en-

codings of visual objects in IT cortex [30], without relying on any explicit external categorical labels. Existing predictive-learning models based on error backpropagation [14] have not demonstrated the development of abstract, categorical representations. Previous work has shown that predictive learning can be a useful method for pretraining networks that are subsequently trained using human-generated labels, but here we focus on the formation of systematic categories *de-novo*.

To determine if our biologically based predictive learning model (Figure b) can naturally form such categorical encodings in the complete absence of external category labels, we showed the model brief movies of 156 3D object exemplars drawn from 20 different basic-level categories (e.g., car, stapler, table lamp, traffic cone, etc.) selected from the CU3D-100 dataset [31]. The objects moved and rotated in 3D space over 8 movie frames, where each frame was sampled at the alpha frequency (Figure c). There were also saccadic eye movements every other frame, with an efferent copy signal to enable full prediction of the effects of the eye movement, which allows the model to capture predictive remapping (a widely-studied signature of predictive learning in the brain) [32,33], and introduces additional predictive-learning challenge. The only learning signal available to the model was the temporal difference prediction error between what it predicted to see in the next frame, compared to what was actually seen.

We performed a representational similarity analysis (RSA) on the learned activity patterns at each layer in the model, and found that the highest IT layer (TE) produced a systematic organization of the 156 3D objects into 5 categories (Figure a), which visually correspond to the overall shape of the objects (pyramid-shaped, vertically-elongated, round, boxy / square, and horizontally-elongated). This organization of the objects matches that produced by humans making shape similarity judgments on the same set of objects, using the V1 reconstruction as shown in Figure c to

capture the model's coarse-grained perception (Figure b; see supplementary information for methods and further analysis). Critically, Figure c shows that the overall similarity structure present in IT layers (TEO, TE) of the biological model is significantly different from the similarity structure at the level of the V1 primary visual input. Thus the model, despite being trained only to generate accurate visual input-level predictions, has learned to represent these objects in an abstract way that goes beyond the raw input-level information. Furthermore, because this abstract category organization reflects the overall visual shapes of the objects as judged by human participants, this suggests that the model is extracting the geometrical shape information that is apparent once these objects are encoded in representations that are invariant to differences in motion, rotation, and scaling present in the V1 visual inputs. We further verified that at the highest IT levels in the model, a consistent, spatially-invariant representation is present across different views of the same object (e.g., the average correlation across frames within an object was .901). This is also evident in Figure a by virtue of the close similarity across multiple objects within the same category.

Further evidence for the progressive nature of representation development in our model is shown in Figure , which compares the similarity structures in layers V4 and IT in macaque monkeys [30] with those in corresponding layers in our model. In both the monkeys and our model, the higher IT layer builds upon and clarifies the noisier structure that is emerging in the earlier V4 layer. Considerable other work has also compared DCNN representations with these same data from monkeys [30], but it is essential to appreciate that those DCNN models were explicitly trained on the category labels, making it somewhat less than surprising that such categorical representations developed. By contrast, we reiterate that our model has discovered its categorical representations entirely on its own, with no explicit categorical inputs or training of any kind.

Figure shows the results from a purely backpropagation-based version of the same model architecture. In this model, the highest layers in the network form a simple binary category structure overall, and the detailed item-level similarity structure does not diverge significantly from that present at the lowest V1 inputs, indicating that it has not formed novel systematic structured representations, in contrast to those formed in the biologically based model. Thus, it is clear that the additional biologically motivated properties of the original model are playing a critical role in the development of abstract categorical representations. These properties include: excitatory bidirectional connections, inhibitory competition, and an additional Hebbian form of learning that serves as a regularizer (similar to weight decay) on top of predictive error-driven learning [12,34].

Each of these properties could promote the formation of categorical representations. Bidirectional connections enable top-down signals to consistently shape lower-level representations, creating significant attractor dynamics that cause the entire network to settle into discrete categorical attractor states. By contrast, backpropagation networks typically lack these kinds of attractor dynamics, and this could contribute significantly to their relative lack of categorical learning. Hebbian learning drives the formation of representations that encode the principal components of activity correlations over time, which can help more categorical representations coalesce (and results below already indicate its importance). Inhibition, especially in combination with Hebbian learning, drives representations to specialize on more specific subsets of the space. Ongoing work is attempting to determine which of these is essential in this case (perhaps all of them) by systematically introducing some of these properties into the backpropagation model, though this is difficult because full bidirectional recurrent activity propagation, which is essential for conveying error signals top-down in the biological network, is incompatible with the standard efficient form of error

backpropagation, and requires much more computationally intensive and unstable forms of fully recurrent backpropagation [35,36]. Furthermore, Hebbian learning requires inhibitory competition which is difficult to incorporate within the backpropagation framework.

Figure shows just a few of the large number of parameter manipulations that have been conducted to develop and test the final architecture. For example, we hypothesized that separating the overall prediction problem between a spatial *Where* vs. non-spatial *What* pathway [37,38], would strongly benefit the formation of more abstract, categorical object representations in the *What* pathway. Specifically, the *Where* pathway can learn relatively quickly to predict the overall spatial trajectory of the object (and anticipate the effects of saccades), and thus effectively regress out that component of the overall prediction error, leaving the residual error concentrated in object feature information, which can train the ventral *What* pathway to develop abstract visual categories. Figure a shows that, indeed, when the *Where* pathway is lesioned, the formation of abstract categorical representations in the intact *What* pathway is significantly impaired. Figure b shows that full predictive learning, as compared to just encoding and decoding the current state (which is much easier computationally, and leads to much better overall accuracy), is also critical for the formation of abstract categorical representations — prediction is a "desirable difficulty" [39]. Finally, Figure c shows the impact of reducing Hebbian learning, which impairs category learning as expected.

In conclusion, we have demonstrated that learning based strictly on predicting what will be seen next is, in conjunction with a number of critical biologically motivated network properties and mechanisms, capable of generating abstract, invariant categorical representations of the overall shapes of objects. The nature of these shape representations closely matches human shape

similarity judgments on the same objects. Thus, predictive learning has the potential to go beyond the surface structure of its inputs, and develop systematic, abstract encodings of the "deeper" structure of the environment. Relative to existing machine-learning-based approaches in "deep learning", which have generally focused on raw categorization accuracy measures using explicit category labels or other human-labeled inputs, the results here suggest that focusing more on the nature of what is learned in the model might provide a valuable alternative approach. Considerable evidence in cognitive neuroscience suggests that the primary function of the many nested ("deep") layers of neural processing in the neocortex is to *simplify* and aggressively *discard* information [40], to produce precisely the kinds of extremely valuable abstractions such as object categories, and, ultimately, symbol-like representations that support high-level cogntive processeses such as reasoning and problem-solving [41,42]. Thus, particularly in the domain of predictive or generative learning, the metric of interest should not be the accuracy of prediction itself (which is indeed notably worse in our biologically based model compared to the DCNN-based PredNet and back-propagation models), but rather whether this learning process results in the formation of simpler, abstract representations of the world that can in turn support higher levels of cognitive function.

Considerable further work remains to be done to more precisely characterize the essential properties of our biologically motivated model necessary to produce this abstract form of learning, and to further explore the full scope of predictive learning across different domains. We strongly suspect that extensive cross-modal predictive learning in real-world environments, including between sensory and motor systems, is a significant factor in infant development and could greatly multiply the opportunities for the formation of higher-order abstract representations that more compactly and systematically capture the structure of the world [43]. Future versions of these mod-

els could thus potentially provide novel insights into the fundamental question of how deep an understanding a pre-verbal human, or a non-verbal primate, can develop [11,44], based on predictive learning mechanisms. This would then represent the foundation upon which language and cultural learning builds, to shape the full extent of human intelligence.

References

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

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

3. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **61**, 85–117 (2015).

4. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).

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

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

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

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

9. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences* **40** (2017/ed).

10. Elman, J. L. Finding Structure In Time. *Cognitive Science* **14**, 179–211 (1990).

11. Elman, J. *et al. Rethinking Innateness: A Connectionist Perspective on Development* (MIT Press, Cambridge, MA, 1996).

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

13. O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E. & Contributors. *Computational Cognitive Neuroscience* (Wiki Book, 1st Edition, URL: http://ccnbook.colorado.edu, 2012).

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

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

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

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

18. Franceschetti, S. *et al.* Ionic mechanisms underlying burst firing in pyramidal neurons: Intracellular study in rat sensorimotor cortex. *Brain Research* **696**, 127–139 (1995).

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

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

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

22. Buffalo, E. A., Fries, P., Landman, R., Buschman, T. J. & Desimone, R. 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**, 11262–11267 (2011).

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

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

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

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

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

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

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

30. Cadieu, C. F. *et al.* Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology* **10**, e1003963 (2014).

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

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

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

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

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

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

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

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

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

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

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

42. O'Reilly, R. C. *et al.* How Limited Systematicity Emerges: A Computational Cognitive Neuroscience Approach. In Calvo, I. P. & Symons, J. (eds.) *The Architecture of Cognition: Rethinking Fodor and Pylyshyn*[1]*s Systematicity Challenge* (MIT Press, Cambridge, MA, 2014).

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

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

Acknowledgements

We thank Dean Wyatte, Tom Hazy, Seth Herd, Kai Krueger, Tim Curran, David Sheinberg, Lew Harvey, Jessica Mollick, Will Chapman, Helene Devillez, and the rest of the CCN Lab for many helpful comments and suggestions.

Funding: Supported by: ONR grants ONR N00014-19-1-2684 / N00014-18-1-2116, N00014-14-1-0670 / N00014-16-1-2128, N00014-18-C-2067, N00014-13-1-0067, D00014-12-C-0638. This work utilized the Janus supercomputer, which is supported by the National Science Foundation (award number CNS-0821794) and the University of Colorado Boulder. The Janus supercomputer is a joint effort of the University of Colorado Boulder, the University of Colorado Denver and the National Center for Atmospheric Research.

Author Contributions: RCO developed the model, performed the non-PredNet simulations, and drafted the paper. JLW performed the PredNet simulations and analysis, and edited the paper. JR contributed to developing the model and edited the paper.

Competing Interests: R. C. O'Reilly is Chief Scientist at eCortex, Inc., which may derive indirect benefit from the work presented here.

Data and Materials Availability: All data and materials will be available at `https://github.com/ccnlab/deep-obj-cat` upon publication.

Supplementary Information

Materials and Methods

Figures S1 - S9

Table S1

Figure 1 **a)** Temporal evolution of information flow in the DeepLeabra model predicting visual sequences, over two alpha cycles of 100 msec each. In each alpha cycle, the V2 Deep layer (lamina 5, 6) uses the prior 100 msec of context to generate a prediction (*minus* phase) on the pulvinar thalamic relay cells (TRC). The bottom-up outcome is driven by V1 5IB strong driver inputs (*plus* phase); error-driven learning occurs as a function of the temporal difference between these phase, in both superficial (lamina 2, 3) and deep, sent via broad pulvinar projections. 5IB bursting in V2 drives update of temporal context in V2 Deep layers, and also the plus phase in higher area TRC, to drive higher-level predictive learning. See supplementary information (SI) for more details. **b)** The three-visual-stream deep predictive learning model (*What-Where-Integration, WWI* model). The dorsal *Where* pathway learns first, using easily-abstracted *spatial blobs*, to predict where an object will move next, based on prior motion history, visual motion, and saccade efferent copy signals. This drives strong top-down inputs to lower areas with accurate spatial predictions, leaving the *residual* error concentrated on *What* and *What * Where* integration information. The V3 and DP (dorsal prelunate) areas constitute the *What * Where* integration pathway, helping bind features and locations. V4, TEO, and TE are the *What* pathway, learning abstracted object category representations, which also drive strong top-down inputs to lower areas. *s* suffix = superficial layer, *d* = deep layer, and *p* = pulvinar. **c)** An example sequence of 8 frames (8 alpha cycles) that the model learned to predict, with the reconstruction of each image based on the V1 gabor filters (*V1 recon*), and a reconstruction of the model-generated prediction for each frame over the higher resolution V1hp pulvinar layer (to compare against V1 recon, correlation value r shown). The relatively low resolution encoding of the image makes these somewhat difficult to interpret, but the r values are well above the r's for each V1 state compared to the previous time step (mean = .38, min of .16 on frame 4 when the prediction is at .57 – see SI for more analysis), indicating that the model has learned somewhat vague but broadly accurate predictions that go beyond e.g.,

just copying the previous time step. The eye icons indicate when a saccade occurred.

Figure 2 **a)** Category similarity structure that developed in the highest layer, TE, of the biologically based predictive learning model, showing *1-correlation* similarity of the TE representation for each 3D object against every other 3D object (156 total objects). Blue cells have high similarity, and model has learned block-diagonal clusters or categories of high-similarity groupings, contrasted against dissimilar off-diagonal other categories. Clustering maximized the overall average *within - between* correlation distance across given set of clusters (see supplementary informations for details). Note that all items from the same "objective" basic-level object categories (N=20) are reliably sorted within a given learned category, so these categories subsume our more fine-grained object categories. **b)** Human similarity ratings for the same 3D objects, presented with the V1 reconstruction (see Fig 1c) to capture coarse perception in our model, aggregated at the level of the 20 basic-level categories since we could not sample the entire 156 x 156 matrix. Each cell is 1 - proportion of time the given pair of objects was rated as more similar than another pair of objects (see supplementary information for details of the experiment). The resulting similarity matrix generally exhibits the same categorical structure as the model (confirmed by permutation testing and agglomorative cluster analysis). **c)** Emergence of abstract category structure over the hierarchy of layers. Red line shows correlation similarity between the similarity matrix for TE (shown in panel a) against the similarity matrix computed for every other layer, and the black line shows the correlation similarity for the V1 layer matrix against every other layer (1 = identical; 0 = orthogonal). Both show that IT layers (TEO, TE) progressively differentiate from raw input similarity structure present in V1, and, critically, that the model has learned structure beyond that present in the input.

Figure 3 Comparison of progression from V4 to IT in macaque monkey visual cortex (top row, from Cadieu et al, 2014) versus same progression in model (replotted using comparable color scale). Although the underlying categories are different, and the monkeys have a much richer

multi-modal experience of the world that could help reinforce categories such as foods and faces, our model nevertheless shows a similar qualitative progression in extent of stronger categorical structure in IT, where the block-diagonal highly similar representations are more consistent across categories, and the off-diagonal differences are stronger and more consistent as well (i.e., categories are also more clearly differentiated). Note that the critical difference in our model versus those compared in Cadieu et al 2014 and related papers is that they explicitly trained their models on category labels, whereas our model is *entirely self-organizing* and has no external categorical training signal.

Figure 4 **a)** Category similarity structure in the highest IT layer (TE) of the backpropagation (Bp) model with the same What / Where structure. Only two broad categories are present, and the lower *max* distance (0.3 vs. 1.5 in biological model) indicates that overall the patterns are highly similar. **b)** Similarity structure for the PredNet model, in the highest of its layers (layer 3), which is even less differentiated (max = 0.15) but overall follows the same broad category structure. **c)** Comparison of similarity structures across layers in the Bp model (compare to Figure 2c): unlike in the biological model, the V1 structure is largely preserved across layers, and is little different from the structure that best fits the TE layer shown in panel a, indicating that the model has not developed abstractions beyond the structure present in the visual input. Layer V3 is most directly influenced by spatial prediction errors in its connections with the dorsal pathway, so it differs from both in strongly encoding position information. **d)** The best fitting V1 structure, which has 2 broad categories and banana is in a third category by itself. The lack of dark blue on the block diagonal indicates that these categories are overall quite weak, and every item is fairly dissimilar from every other. **e)** The same similarities shown in panel a for Bp TE also fit reasonably well in the V1 structure (and they have a similar average within - between contrast differneces, of -0.0838 and -0.0513). **f)** The similarity structure from the biological model does *not* fit well within the V1

organization (the blue is not aligned along the block diagonal, and the yellow is not strictly off-diagonal), consistent with the large difference in average contrast distance (0.5071 for the best categories vs. 0.3070 for the V1 categories).

Figure 5 Effects of various manipulations on the extent to which the TE layer representations differentiate from the V1 structure. *Std* is the same result shown in Figure 2c from the intact model (see caption there for further explanation), and all manipulations significantly impair the development of abstract TE categorical representations (i.e., the TE representations are more similar to V1 and the other layers). **a)** Dorsal *Where* pathway lesions, including lateral inferior parietal sulcus (LIP), V3, and dorsal prelunate (DP). This pathway is essential for regressing out location-based prediction errors, so that the residual errors concentrate feature-encoding errors that train the *What* pathway. **b)** Allowing the deep layers full access to current-time information, thus effectively eliminating the prediction demand and turning the network into an auto-encoder, which significantly impairs representation development, and supports the importance of the challenge of predictive learning for developing deeper, more abstract representations. **c)** Reducing the strength of Hebbian learning by 20% (from 2.5 to 2), demonstrating the essential role played by this form of learning on shaping categorical representations. Eliminating Hebbian learning entirely prevented the model from learning anything at all, as it also plays a critical regularization and shaping role on learning.