A stimulus-computable proxy for the primate visual system overturns evidence against perirhinal involvement in perception

Tyler Bonnen *[a] and Mark A.G. Eldridge[b]

[a]Stanford University, Stanford CA
[b]National Institute of Mental Health, Bethesda MD

Animals seamlessly integrate sensory activity with previously encountered, behaviorally relevant experience. Neuroanatomical structures within the medial temporal lobe, such as perirhinal cortex (PRC), are known to enable these memory-related behaviors. Yet there is an enduring debate over PRC involvement in perception, beset by decades of seemingly inconsistent experimental outcomes. Here, we leverage a computational approach able to predict choice behavior directly from experimental stimuli, using a biological plausible approximation of the primate ventral visual stream (VVS). We deploy this approach on visual experiments used as evidence against PRC involvement in perception. Our VVS model predicts behavior with striking precision, for both PRC-lesioned and -intact subjects, suggesting that these stimuli are not diagnostic of PRC involvement in perception. Our results suggest that *apparent* inconsistencies in the literature are due to reliance on informal, descriptive accounts of experimental properties; stimulus-computable models of perception offer an alternative.

1 Introduction

Neuroanatomical structures within the medial temporal lobe (MTL) are known to support memory-related behaviors (Eichenbaum and Cohen, 2004; Scoville and Milner, 1957). Yet there is an enduring debate over their involvement in perception (Bussey et al., 2003; Suzuki, 2009). In the object perception literature, this debate has centered on perirhinal cortex (PRC), an MTL structure situated at the apex high-level sensory cortices (Miyashita, 2019). Decades of causal (lesion), correlational (neuroimaging), and computational studies in rodents, human-, and non-human primates, have resulted in a pattern of seemingly inconsistent experimental findings—some studies demonstrating PRC-related perceptual deficits (e.g. Murray and Bussey, 1999; Barense et al., 2007; Bussey et al., 2002; Inhoff et al., 2019; Lee et al., 2006; Lee et al., 2005), while others appear to reveal no such impairments (Buffalo et al., 1998a; Buffalo et al., 1998b; Knutson et al., 2012; Stark and Squire, 2000). Identifying PRC-dependent perceptual processing would require experimentalists to isolate PRC functions from those of surrounding, perceptual cortices. Unfortunately, experimentalists have been forced to rely on informal, descriptive accounts of stimulus properties (e.g. 'feature ambiguity') in order to identify stimulus properties that uniquely depend PRC.

Here we employ a computational approach that formalizes visual perceptual demands directly from experimental stimuli: a task optimized convolutional neural network, validated on electrophysiological recordings throughout the primate ventral visual stream (VVS). We employ this modeling approach to evaluate visual discrimination experiments administered to PRC-intact and -lesioned non-human primates. These experiments have been designed to evaluate PRC-involvement in perception, using stimuli with high 'feature ambiguity,' which has been shown to depend on PRC (Murray and Bussey, 1999). However, in all experiments, we find that this computational proxy for the VVS is able to predict the behavior of PRC-intact and -lesioned subjects with striking precision. We perform this analysis using metrics employed by the original authors, as well as more fine-grained subject- and image-level analyses. Our results suggest that a linear readout of the VVS is sufficient for performance on these stimulus sets. That is, while these stimuli have been used to evaluate PRC involvement in perception—and, subsequently, to claim PRC is not involved in perception—our results suggest these experiments are not diagnostic.

2 Results

We analyze stimuli and behavioral data from an ostensibly 'complex' stimuli set administered to PRC-lesioned and -intact subjects (Eldridge et al., 2018). Across four experiments in this study, stimuli are composed of cats, dogs, and 'morphed' images that parametrically vary the percent of category-relevant information present in each trial. These 'morphed' stimuli were designed to evaluate PRC involvement in perception by creating maximal 'feature ambiguity,' a perceptual

*Corresponding author's email: bonnen@stanford.edu

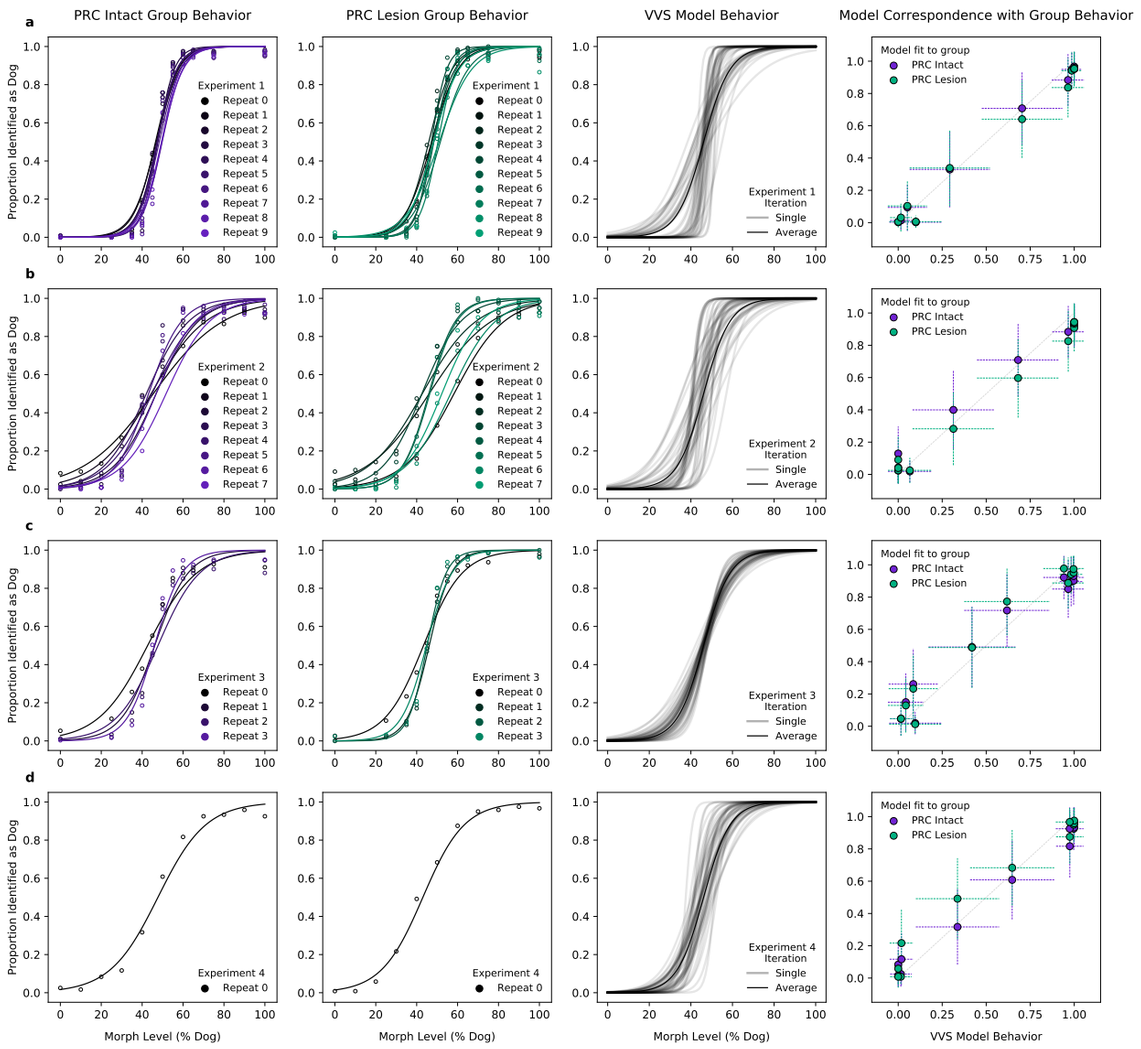

Figure 1: **VVS model predicts PRC-intact and -lesioned behaviors.** Stimuli have been designed to evaluate the role of PRC involvement in perception, as they are thought to require putatively 'complex' perceptual representations not supported by canonical visual cortices. PRC-intact (purple, n=3) and -lesioned (green, n=3) subjects exhibit a similar pattern of responses within and across experiments **(a-d)**. This has been taken as evidence against PRC involvement in perception by Eldridge et al., 2018. To evaluate this claim, we present experimental stimuli to a computational proxy for the ventral visual stream (VVS) and extract model responses from a layer that corresponds with 'high-level' perceptual (inferior temporal: IT) cortex. Using model responses from an 'IT-like' layer, we learn to predict the category membership of each stimulus, testing this linear mapping in left-out images across multiple train-test iterations (black). We evaluate correspondence between model behavior and PRC-intact and -lesioned subjects (far right). This computational proxy for the VVS predicts the choice behavior of PRC-intact (purple) and -lesioned (green) grouped subjects (error bars indicate standard deviation from the mean, across model iterations and subject choice behaviors). A linear readout of the VVS appears to be sufficient to perform these tasks, suggesting no need for PRC-involvement. Our results suggest that this stimulus set is not diagnostic of PRC involvement in perception, overturning previous claims against PRC involvement in perception using these very same experimental stimuli.

quality purported to leads to PRC dependence. For example, '10% morphs' are 90% cat and 10% dog. Each trial requires that subjects perform a binary decisions as to the category membership of the stimulus, judging whether an image was more 'dog-like' or 'cat-like.' After each trial, across all morph levels, subjects are rewarded for responses that correctly identify which category best fits the image presented (e.g. 10%='cat', 80%='dog'). Here we evaluate data from two groups of monkeys: an unoperated control group (n=3) and a group with bilateral removal of rhinal cortex (including peri- and ento-rhinal cortex). We formulate the modeling problem as a binary forced choice (i.e. 'dog'=1, 'cat'=0), presenting the model with experimental stimuli, optimizing a linear mapping from model responses to predict the category label, then evaluating this linear mapping in independent data.

First, we estimate model correspondence to PRC-lesioned and -intact subjects with the aggregate metrics used by the original authors; averaging performance on all images within each

morph level (e.g. 10%, 20%, etc.) across all subjects within each group (Fig. 1; PRC-intact (purple) and -lesioned (green) subjects across experiments (a-d). As reported in Eldridge et al., 2018, there is not a significant difference between the choice behaviors of PRC-lesioned and -intact subjects (lesion group fails to predict variance in choice behavior: $R^2 = 0.00$ $\beta = -0.01$, $F(1, 32) = -0.10, P = 0.925$). A computational proxy for the VVS exhibits the same qualitative pattern of behavior as both groups (Fig. 1, model performance across multiple train-test iterations in black). Comparing these model predictions to each group (Fig. 1, far right) we observe a striking correspondence between model performance with PRC-intact (purple: $R^2 = 0.98$ $\beta = 0.97$, $t(21) = 33.12, P = 6$ x 10^{-19}) and -lesioned subjects (green: $R^2 = 0.99$ $\beta = 0.96$, $t(21) = 57.38, P = 1$ x 10^{-23}). Focusing on individual subject behavior, we again observe the same relationship for both PRC-intact (e.g. subject 0; $R^2 = 0.99$ $\beta = 1.01$, $t(21) = 39.30, P = 2$ x 10^{-20}) and -lesioned (e.g. subject 4: $R^2 = 0.99$ $\beta = 1.01$, $t(21) = 45.01, P = 1$ x 10^{-21}) groups. Thus, our computational proxy for the VVS recovers both the group- and subject-level behavior of PRC-intact and -lesioned performance reported in Eldridge et al., 2018.

This analysis only offers a coarse summary of PRC-lesioned and -intact behavior. We next determine whether this modeling approach is able to predict more fine-grained choice behavior. For each subject, we first estimate the split-half reliability of choice behavior across multiple presentations of the same image, using the two experiments with sufficient presentations of each image to perform this analysis (experiment one and two). Using R^2 as a measure of fit across split-halves within each experiment, we find highly consistent within-subject reliable estimates for both PRC-intact (e.g. median $R^2_{\text{exp1}} = .94$, median $R^2_{\text{exp2}} = .86$) and -lesioned (e.g. median $R^2_{\text{exp1}} = .91$, median $R^2_{\text{exp2}} = .90$) subjects. Similarly, we observe highly reliable image-level choice behaviors between subjects (e.g. median $R^2_{\text{exp1}} = .86$, median $R^2_{\text{exp2}} = .79$). The within-subject reliability estimates are significantly greater that those between subjects (unpaired ttest $P = 0.001$). These results suggest that there is meaningful within- and between-subject variance in the image-by-image choice behaviors of experimental subjects (2a: PRC-intact subjects purple, PRC-lesioned subjects green; between-group reliability in grey; within-subject reliability presented also in 2b).

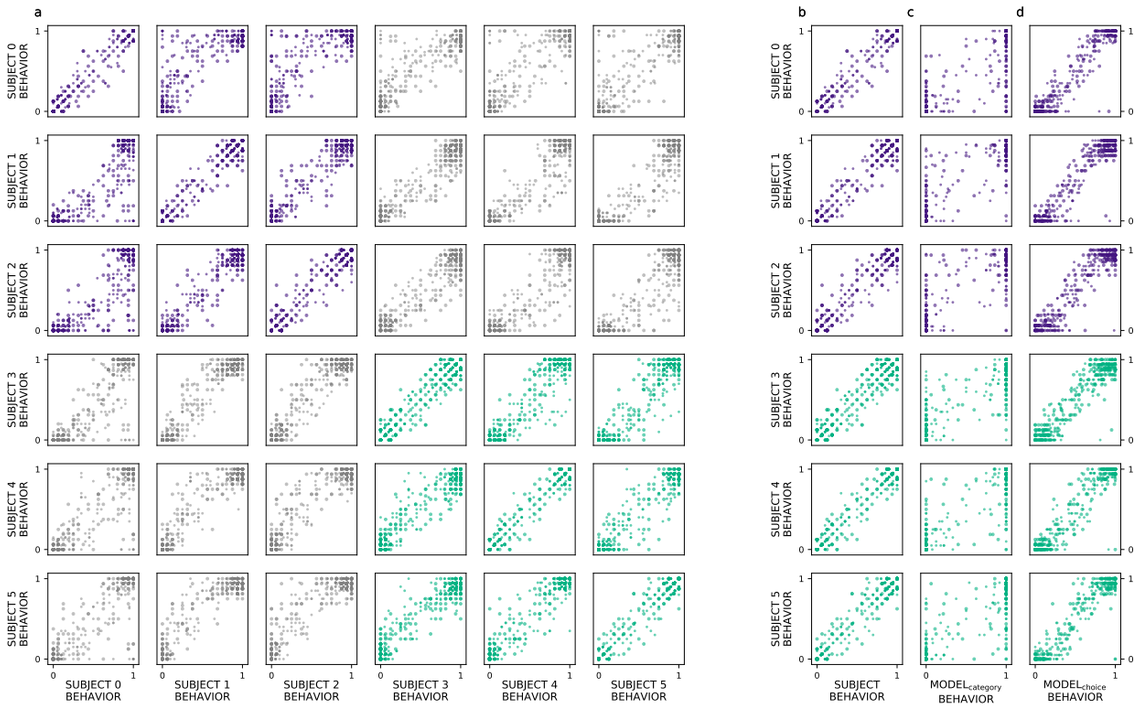

Figure 2: **Image-level correspondence between VVS model and subject choice behavior.** We estimate the within subject reliability over 100 split-halves (on the diagonal) for both PRC-intact (purple, n=3) and -lesioned (green, n=3) choice behavior **(a)**. We estimate the between-subject reliability by comparing the averaged image-level behavior between subjects (off-diagonal), both between PRC-intact subjects (purple), between PRC-lesioned subjects (green), and between lesioned groups (grey). For visualization purposes, we isolate the within-subject split-half reliability estimates across all subjects in **(b)**. While the model optimized to predict *category membership* (i.e. trained on the ground truth labels, Model$_{\text{category}}$) predicts the aggregate metrics in Fig. 1, at the resolution of single images model and subject choice behaviors are markedly different **(c)**. If, instead of using the category membership as a training objective, we instead build models to estimate *choice behavior* of individual subjects (Model$_{\text{choice}}$), we are able to fit the image-level choice behaviors of each subject.

The same model reported in previous analysis (Fig. 1c) is able to significantly predict image-level choice behaviors for both PRC-lesioned (e.g. subject 3: $R^2 = 0.86$ $\beta = 0.81$, $F(1, 438) = 52.79, P = 5 \text{ x } 10^{-192}$) and -intact subjects (e.g. subject 1: $R^2 = 0.87$ $\beta = 0.88$, $F(1, 438) = 53.24, P = 2 \text{ x } 10^{-193}$). However, this is to be expected of any model that accurately performs these tasks, given the near-ceiling performance of experimental subjects on these visual discrimination tasks. To determine whether these behaviors are 'subject-like', we impose a more stringent test, evaluating the model behavior in relation to distributions of within- and between-subject reliability estimates. More concretely, we determine the likelihood of observing the model fits to subject behavior under the observed distributions of within- and between-subject reliability, again using R^2 as a measure of fit. We determine the likelihood of observing the model fits to each subject's behavior under the empirical distribution (across train-test splits, n=100; median $P(\text{model} \mid \text{reliability}_{\text{between-subject}}) = 0.024$). That is, while the model is predicting the aggregate performance across images, it is notably—and significantly—different than the distribution of between-subject variation observed in this dataset. This relationship is also evident when visually comparing model behavior alongside subject behavior (Fig. 2c).

While the preceding analysis diverges with image-level choice behaviors, simple modifications to this modeling approach enable us to predict these fine-grained choice behaviors: instead of training a linear readout of model responses to choose the *correct* category labels, here we take the *choice behavior* of experimental subjects as our learning objective. We learn a weighted mapping from model responses to the choice behavior of each experimental subject, using the same cross-validated approach use to predict ground truth labels, simply changing the training objective. Again, this choice-trained model ($\text{model}_{\text{choice}}$) significantly predicts the performance of PRC-intact (e.g. subject1: $R^2 = 0.94$ $\beta = 0.01$, $F(1, 438) = 84.75, P = 8 \text{ x } 10^{-274}$) and lesioned subjects (e.g. subject4 $R^2 = 0.94$ $\beta = 0.01$, $F(1, 438) = 81.19, P = 4 \text{ x } 10^{-266}$). However, the choice-trained model is a significantly better match to image-level behavior than models trained on ground truth labels (wilcoxon rank-sum test between $\text{model}_{\text{choice}}$ and $\text{model}_{\text{category}}$: median $P = 2 \text{ x } 10^{-15}$) and is many times more likely to be observed within the true distribution of between-subject reliability estimates ($P(\text{model}_{\text{choice}} | \text{reliability}_{\text{between-subject}}) = 0.364$, non-parametric Bayes Factor comparing the likelihood of observing $\text{model}_{\text{choice}}$ over $\text{model}_{\text{category}} = 8.5$). Thus, with simple modifications to the model's training objective, we are able to capture much of the image-level subject choice behavior (Fig. 2d).

3 Discussion

Using a stimulus-computable approximation of the primate VVS, we have evaluated the perceptual demands imposed from multiple experiments originally presented in Eldridge et al., 2018. The original authors concluded that PRC is not involved in perception, given the absence of any PRC-lesioned deficits on these ostensibly 'complex' stimulus sets. Our computational results challenge this claim: We find a striking correspondence between a computational proxy for the VVS and subject behavior, both PRC-intact and -lesioned. We observe this correspondence across multiple resolutions, using the group- and 'morph-' level analyses employed by the original authors, as well as more fined-grained, subject-by-image-level analysis. Our results suggest that these stimuli are not diagnostic of PRC involvement in perception. This should not be interpreted as evidence *supporting* PRC involvement in perception: we can simply state that these stimulus sets are uninformative as to the nature of PRC involvement in perception.

As has been previously reported (Rajalingham et al., 2018), we find that increasingly granular (e.g. image-level) analyses highlight the divergence between computational proxies for the VVS and animal behavior. This observed divergence could have been a consequence of relatively independent modeling components—training on biologically implausible datasets, using improper learning rules, or any number of differences between biological and artificial neural networks (Zhuang et al., 2021). However, we find that simple changes to the model's training objective can generate image-level choice behaviors remarkably consistent with both PRC-lesioned and -intact subjects; while the model responses to incoming stimuli remained unchanged, we were able to learn a readout that better matched subject behavior. Given the degree to which animals are able to flexibly deploy their visual systems, future work isolating PRC-dependent processing from VVS-supported behaviors must ensure that VVS-model failures are not simply a consequence of linear (DiCarlo and Cox, 2007) readout types.

Taken together, this work suggests that apparent inconsistencies in the experimental literature may be due to reliance on informal, descriptive accounts of perceptual demands. While possibility has been suggested (Barense et al., 2007), we have been able to demonstrate the lack of correspondence between descriptive terms and more objective measures through our analysis of Eldridge et al., 2018. We note here that attributes used to design these stimulus sets (i.e. 'feature ambiguity') were well-documented in the PRC-related perceptual literature (Lee et al., 2005; Murray et al.,

2007). However, these terms are informal and underconstrained; it should be no surprise that these terms do not reliably characterize neural function. Beyond offering a case study surrounding the limitations of using informal, descriptive accounts of experimental variables to guide hypothesis evaluation in the neurosciences, we have demonstrated the utility of an alternative approach: using an extensible, computable framework that enables us to move beyond informal, descriptive accounts of experimental properties and evaluate perceptual demands on PRC directly from experimental stimuli.

References

Barense, M. D., Gaffan, D., & Graham, K. S. (2007). The human medial temporal lobe processes online representations of complex objects. *Neuropsychologia*, *45*(13), 2963–2974.

Buffalo, E. A., Reber, P. J., & Squire, L. R. (1998a). The human perirhinal cortex and recognition memory. *Hippocampus*, *8*(4), 330–339.

Buffalo, E. A., Stefanacci, L., Squire, L. R., & Zola, S. M. (1998b). A reexamination of the concurrent discrimination learning task: The importance of anterior inferotemporal cortex, area te. *Behavioral neuroscience*, *112*(1), 3.

Bussey, T. J., Saksida, L. M., & Murray, E. A. (2002). Perirhinal cortex resolves feature ambiguity in complex visual discriminations. *European Journal of Neuroscience*, *15*(2), 365–374.

Bussey, T. J., Saksida, L. M., & Murray, E. A. (2003). Impairments in visual discrimination after perirhinal cortex lesions: Testing 'declarative' vs. 'perceptual-mnemonic' views of perirhinal cortex function. *European Journal of Neuroscience*, *17*(3), 649–660.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in cognitive sciences*, *11*(8), 333–341.

Eichenbaum, H., & Cohen, N. J. (2004). *From conditioning to conscious recollection: Memory systems of the brain*. Oxford University Press on Demand.

Eldridge, M. A., Matsumoto, N., Jnr, J. H. W., Masseau, E. C., Saunders, R. C., & Richmond, B. J. (2018). Perceptual processing in the ventral visual stream requires area te but not rhinal cortex. *Elife*, *7*, e36310.

Inhoff, M. C., Heusser, A. C., Tambini, A., Martin, C. B., O'Neil, E. B., Köhler, S., Meager, M. R., Blackmon, K., Vazquez, B., Devinsky, O., et al. (2019). Understanding perirhinal contributions to perception and memory: Evidence through the lens of selective perirhinal damage. *Neuropsychologia*, *124*, 9–18.

Knutson, A. R., Hopkins, R. O., & Squire, L. R. (2012). Visual discrimination performance, memory, and medial temporal lobe function. *Proceedings of the National Academy of Sciences*, *109*(32), 13106–13111.

Lee, A. C., Buckley, M. J., Gaffan, D., Emery, T., Hodges, J. R., & Graham, K. S. (2006). Differentiating the roles of the hippocampus and perirhinal cortex in processes beyond long-term declarative memory: A double dissociation in dementia. *Journal of Neuroscience*, *26*(19), 5198–5203.

Lee, A. C., Bussey, T. J., Murray, E. A., Saksida, L. M., Epstein, R. A., Kapur, N., Hodges, J. R., & Graham, K. S. (2005). Perceptual deficits in amnesia: Challenging the medial temporal lobe 'mnemonic'view. *Neuropsychologia*, *43*(1), 1–11.

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. *Journal of Neuroscience*, *35*(39), 13402–13418.

Miyashita, Y. (2019). Perirhinal circuits for memory processing. *Nature Reviews Neuroscience*, *20*(10), 577–592.

Murray, E. A., & Bussey, T. J. (1999). Perceptual–mnemonic functions of the perirhinal cortex. *Trends in cognitive sciences*, *3*(4), 142–151.

Murray, E. A., Bussey, T. J., & Saksida, L. M. (2007). Visual perception and memory: A new view of medial temporal lobe function in primates and rodents. *Annu. Rev. Neurosci.*, *30*, 99–122.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*(33), 7255–7269.

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry*, *20*(1), 11.

Stark, C. E., & Squire, L. R. (2000). Intact visual perceptual discrimination in humans in the absence of perirhinal cortex. *Learning & Memory*, *7*(5), 273–278.

Suzuki, W. A. (2009). Perception and the medial temporal lobe: Evaluating the current evidence. *Neuron*, *61*(5), 657–666.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3).

4 Methods

4.1 Model fit to electrophysiological recordings throughout the VVS

Our modeling approach begins with a task-optimized convolutional neural network, pre-trained to perform object classification on a large-scale dataset (Deng et al., 2009). Using previously collected electrophysiological responses from macaque VVS (Majaj et al., 2015), we identify a model layers that best fit high-level visual cortex: given images as input, we learn a linear mapping between model responses and a single electrode's responses, then evaluate this mapping using independent data (Methods: Model Fit to Electrophysiological Data). For each model layer, this analysis yields a median cross-validated fit to noise-corrected neural responses, for both V4 and IT—corresponding to earlier and later stages of processing within the VVS. As is consistently reported (Rajalingham et al., 2018; Schrimpf et al., 2020; Yamins et al., 2014), early model layers (i.e. first half of layers) better predict neural responses in V4 than do later layers ($t(8) = 2.70, P = .015$), while later layers better predict neural responses in IT, a higher-level region, ($t(8) = 3.70, P = .002$), such that differential fit to IT cortex (IT-V4) increases in with model depth ($\beta = .98$, $F(1, 17) = 20.91$, $P = 10^{-13}$).