Formalizing medial temporal lobe involvement in perception: Making predictions from pixels, not words

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 formalize two competing theories of PRC involvement in visual object perception, situating behavior in relation to the performance supported by the primate ventral visual stream (VVS). Leveraging electrophysiological recordings from high-level visual cortex, we demonstrate that a computational model enables us to estimate VVS-supported performance, directly from experimental stimuli. With this modeling approach, we reevaluate results from visual discrimination experiments administered to PRC-lesioned and -intact macaques. While these experiments have been used as evidence *against* PRC involvement in perception, our modeling results overturn this interpretation: We are able to predict PRC-intact and -lesioned behavior with striking precision, suggesting that perceptual processing beyond the VVS should not be necessary—that is, these stimuli are not diagnostic of PRC involvement in perception. These results suggest that stimulus-computable models have a critical role to play in understanding MTL involvement in primate perceptual behaviors.

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 of high-level sensory cortices (Fig. 1a; Miyashita, 2019). Numerous studies have reported perceptual impairments following lesions to PRC in humans and other animals (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). However, related studies claim that these PRC-related impairments are due to memory-related task demands, or concurrent damage in PRC-adjacent sensory cortex (e.g. Buffalo et al., 1998a; Buffalo et al., 1998b; Knutson et al., 2012; Stark and Squire, 2000). While decades of experimental evidence have been used to evaluate PRC involvement in perception, there have not been objective means of evaluating these competing claims.

A central challenge in this literature has been isolating PRC-dependent behaviors from those supported by PRC-adjacent perceptual cortex. For visual object perception, in the primate, this requires isolating PRC-dependent behaviors from visual discrimination behaviors that depend on the ventral visual stream (DiCarlo and Cox, 2007; DiCarlo et al., 2012). The perceptual-mnemonic hypothesis predicts that PRC-lesioned subjects show perceptual impairments only on stimuli that are sufficiently 'complex'–i.e. not supported by the VVS alone (Murray and Baxter, 2006; Murray and Bussey, 1999; Murray and Wise, 2012). While this prediction is straightforward, interpreting the deficits—or lack thereof—following PRC lesions has required that experimentalists rely on informal, descriptive accounts of perceptual demands. For example, the absence of PRC-related deficits on a given experiment (e.g. in Stark and Squire, 2000) has led some to conclude that these data are evidence against PRC involvement in perception (as per Suzuki and Baxter, 2009), while others have argued that the stimuli in this experiment are not not sufficiently 'complex,' and so are not diagnostic of PRC involvement in perception (as per Bussey and Saksida, 2002). That is, there is not standard operationalization for these descriptive terms used to describe PRC-dependence (e.g. 'complexity,' or 'feature ambiguity').

Here we formalize competing claims surrounding PRC involvement in perception. The perceptual-mnemonic hypothesis predicts that PRC-intact subjects are able to outperform a linear readout of the VVS, but only on tasks that require perceptual representations not supported by the VVS alone (Murray and Bussey, 1999). In line with this hypothesis, PRC-intact human behavior has been shown to outperform a linear readout of electrophysiological recordings from high-level visual cortex (Fig 1c: PRC-intact performance above diagonal, unpaired ttest: $\beta = .24$, $t(31) = 9.50$, $P = 1\text{x}10^{-10}$; originally reported in Bonnen et al., 2020). The perceptual-mnemonic hypothesis predicts that PRC-lesioned behavior will be impaired on these 'complex' stimuli, with PRC-lesioned behavior corresponding not to chance performance, but the performance supported by a linear readout of the VVS (schematized in 1d: left). However, it is possible that this supra-VVS performance is

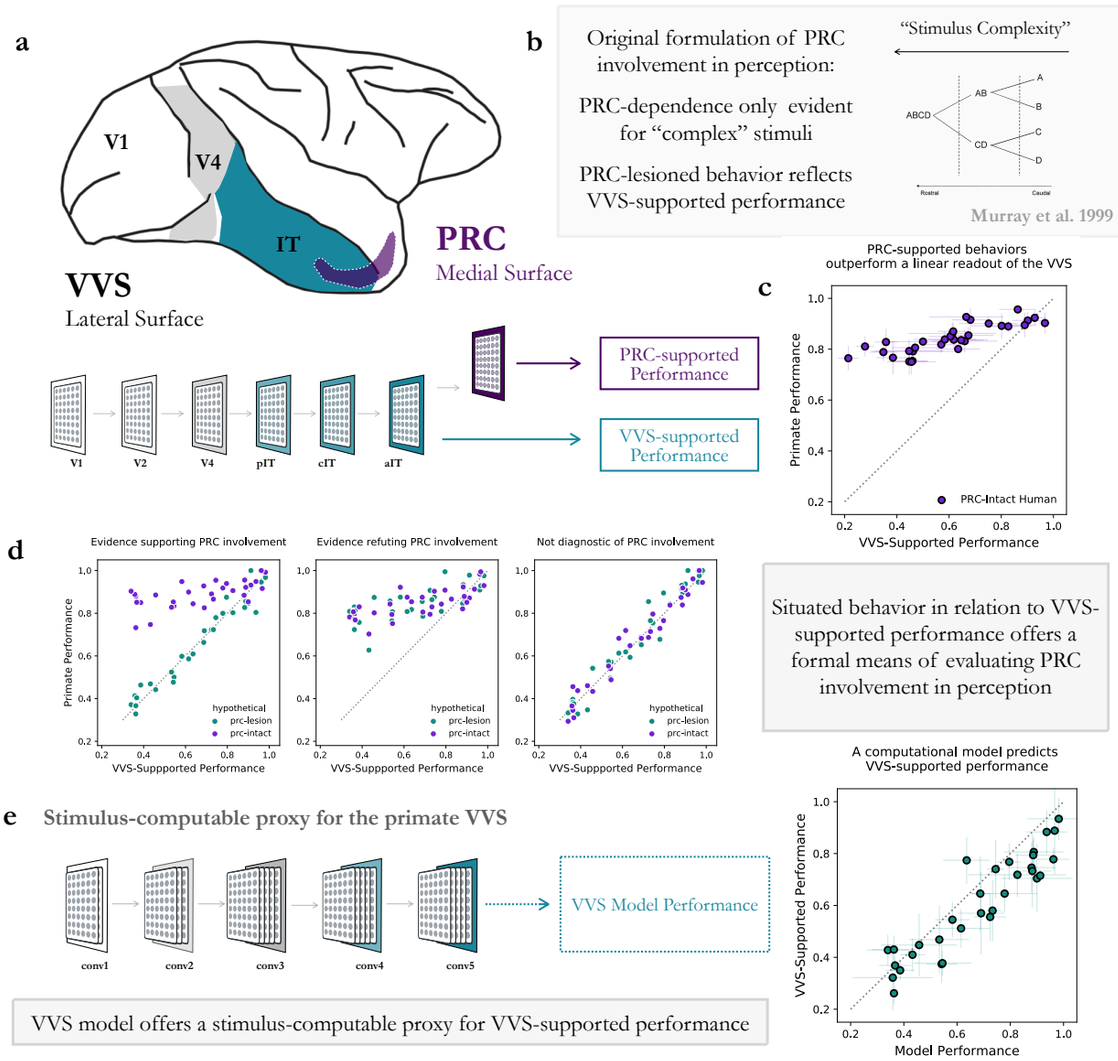

Figure 1: **A computational framework to formalize PRC involvement in perception. (a)** Perirhinal cortex (PRC) is a medial temporal lobe (MTL) structure situated at the apex of the primate ventral visual stream (VVS). **(b)** A perceptual-mnemonic account suggests that, in addition to it's mnemonic functions, PRC enables 'complex' perceptual representations, not supported by canonical sensory cortices alone. Accordingly, PRC-related perceptual deficits are only expected for stimuli that require sufficiently 'complex' representations. **(c)** In agreement with the perceptual-mnemonic hypothesis, it has been shown that PRC-intact subjects outperform a direct readout from high-level visual cortex. Error bars represent standard deviation, for both PRC-intact human performance (y axis) and the behavior supported from a linear readout of electrophysiological recordings collected from inferior temporal (IT) cortex (x axis). **(d)** Given this, we illustrate the patterns of evidence that must be observed within a given experiment in order to support (left) or refute (center) PRC involvement in perception. If, however, a visual discrimination task is supported by the VVS, then this stimulus set can not evaluate—i.e. is not diagnostic—of PRC involvement in perception (right). **(e)** In order to evaluate VVS-supported performance on any arbitrary stimulus set, even in the absence of electrophysiological recordings, we leverage a computational model able to make predictions about VVS-supported performance, directly from experimental stimuli: task optimized convolutional neural networks (schematized, left). We validate the correspondence between this VVS model and electrophysiological recordings collected from high-level visual cortex (right).

supported by neural structures independent of PRC—consistent with a strictly mnemonic interpretation of PRC (schematized in 1d: middle). We note that a linear readout of the VVS is sufficient for many visual behaviors (DiCarlo and Cox, 2007); for example, electrophysiological recordings from primate inferior temporal (IT) cortex predict human-level object recognition behaviors (Majaj et al., 2015). As such, stimulus sets for which VVS-supported performance predicts PRC-intact behavior are not able to evaluate PRC involvement in perception—i.e. are non-diagnostic—as perceptual processing beyond the VVS should not be necessary (schematized in 1d: right).

Operationalizing the perceptual-mnemonic hypothesis, here we assume that VVS-supported performance can serve as a null model for PRC involvement in perception: accepting that PRC is involved in perception requires that PRC-intact subjects outperform VVS-supported performance

(i.e. reject the null) while PRC-lesioned subject behavior reverts to it. To evaluate this hypothesis we situate electrophysiological, lesion, and behavioral data within a shared computational framework, using a computational proxy for the primate VVS. Critically, our approach makes predictions for animal behavior directly from experimental stimuli, obviating the need for descriptive terms like stimulus 'complexity,' instead relying on the model's correspondence with the performance supported by the primate VVS. With this approach, here we model experiments that have previously been administered to PRC-lesioned and -intact subjects—data that has been used as evidence against PRC involvement in perception. This enables us to situate PRC-lesioned and -intact subject behavior in relation to VVS-supported performance, so that we might evaluate—and falsify—competing claims about the relative involvement of PRC visual discrimination behaviors.

2 Results

Here we leverage a stimulus-computable proxy for the primate VVS, enabling us to estimate VVS-supported performance directly from experimental stimuli. We begin with a task-optimized convolutional neural network, pre-trained to perform object classification on a large-scale dataset (Deng et al., 2009); currently, these are the most competitive models of the primate VVS (Rajalingham et al., 2018; Schrimpf et al., 2020; Yamins et al., 2014). We report results from a single instance of this model class (Simonyan and Zisserman, 2014), but note that there results are consistent across all models evaluated (e.g. He et al., 2016). We validate the correspondence between this model class with electrophysiological recordings previously collected from IT cortex (Majaj et al., 2015) in two ways. First, we determine 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, then evaluate this mapping using independent data (Methods: Model fit to electrophysiological data). Additionally, we determine VVS model performance in relation to a linear readout of high-level visual cortex (Methods: Model correspondence with VVS-supported performance). We observe a striking correspondence between VVS model performance and VVS-supported performance (Fig. 1e: $R^2 = 0.86$ $\beta = 0.81$, $F(1, 30) = 13.33, P = 4$ x 10^{-14}). Taken together, these results suggest that this model is a suitable approximation for VVS-supported performance.

With this computational proxy for the VVS, we estimate VVS-supported performance on an ostensibly 'complex' stimulus set administered to PRC-lesioned and -intact subjects. Across four experiments in this study (Eldridge et al., 2018), stimuli are composed of cats, dogs, and 'morphed' images that parametrically vary the percent of category-relevant information present in each trial. For example, '10% morphs' are 90% cat and 10% dog. These 'morphed' stimuli were designed to evaluate PRC involvement in perception by creating maximal 'feature ambiguity,' a perceptual quality purported to elicit PRC dependence in previous work (Bussey et al., 2002, 2006; Murray and Richmond, 2001; Norman and Eacott, 2004). 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.' Within each trial, 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, and learn a linear mapping from model responses to predict the category label. For all analyses, we report the results from then evaluating this linear mapping in independent (i.e. left out) data.

We estimate model correspondence to PRC-lesioned and -intact subjects with the aggregate metrics used by the original authors. Within each experiment (Fig. 2a-d), we average performance on all images within each morph level (e.g. 10%, 20%, etc., on the x axis), aggregating across subjects in each lesion group (Fig. 2; PRC-intact (purple) and -lesioned (green) subjects). As reported in Eldridge et al., 2018, there is not a significant difference between the choice behaviors of PRC-lesioned and -intact subjects (no significant difference between PRC-intact/-lesion groups: $R^2 = 0.00$ $\beta = 0.01$, $F(1, 86) = 0.07, P = 0.941$). A computational proxy for the VVS exhibits the same qualitative pattern of behavior as both groups (Fig. 2, model performance across multiple train-test iterations in black). Moreover, we observe a striking correspondence between between model and PRC-intact behavior (Fig. 2, far right; purple: $R^2 = 0.98$ $\beta = 0.97$, $t(21) = 33.12, P = 6$ x 10^{-19}) as well as -lesioned subjects (green: $R^2 = 0.99$ $\beta = 0.96$, $t(21) = 57.38, P = 1$ x 10^{-23}). Indeed, if we use the same test used to claim no significant difference between PRC-lesion/-intact performance, we do not find a significant difference between primate and model behavior (no significant difference between VVS-model/primate groups: $R^2 = 0.00$ $\beta = -0.01$, $F(1, 86) = -0.11, P = 0.915$).

Given the correspondence between model behavior with this aggregate metrics of PRC-lesioned and -intact performance, we next ask whether we are able to predict more granular behaviors. We introduce a series of split-half reliability analyses that enable us to evaluate how consistent subject

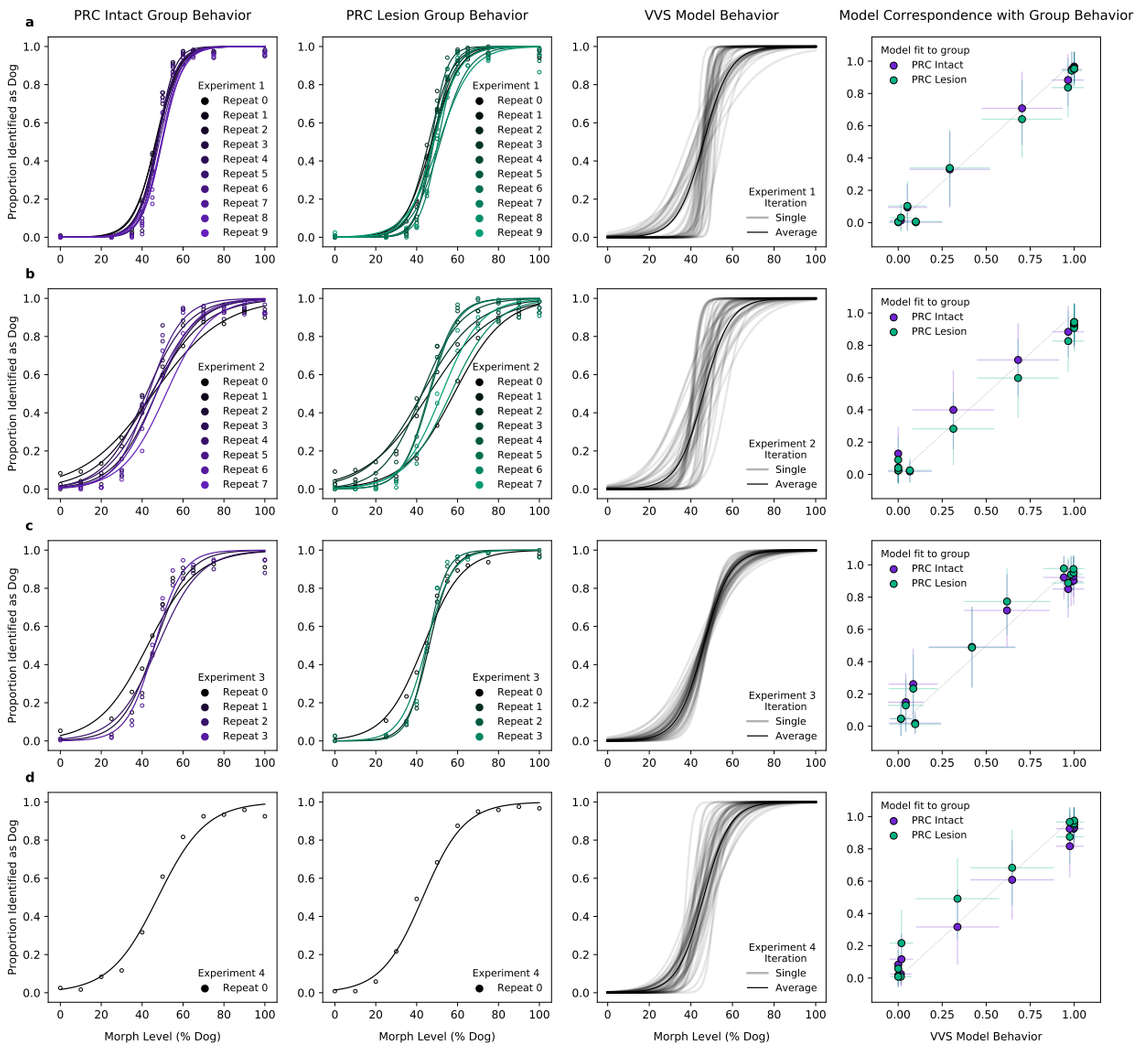

Figure 2: **A computational proxy for the VVS predicts PRC-intact and -lesioned behavior.** Eldridge et al., 2018 designed a visual discrimination task that requires subjects to make 'ambiguous' visual discriminations—a perceptual property argued to depend on PRC (see Supplemental Figure S1 for example stimuli and experimental protocol). However, PRC-intact (purple, n=3) and -lesioned (green, n=3) subjects exhibit a similar pattern of responses within and across these experiments **(a-d)**. This has been taken as evidence against PRC involvement in perception. 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 on 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). As such, 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.

choice behavior is at the level of individual images (Methods: Split-half reliability estimates). Across both experiments, we find consistent image-level choice behaviors 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 (3a: within subject reliability on the diagonal; PRC-intact subjects in purple, PRC-lesioned subjects in green). Between subjects, again, we observe consistent image-level choice behaviors (e.g. median $R^2_{\text{exp1}} = .86$, median $R^2_{\text{exp2}} = .79$). These results indicate there is meaningful within- and between-subject variance in the image-by-image choice behaviors of experimental subjects (Fig. 3a: PRC-intact subjects purple, PRC-lesioned subjects green; between-group reliability in grey), suggesting that this behavior is a suitable target to evaluate our how well we approximate subject behaviors.

First, we report that our computational approach is able to predict subject-level choice behavior

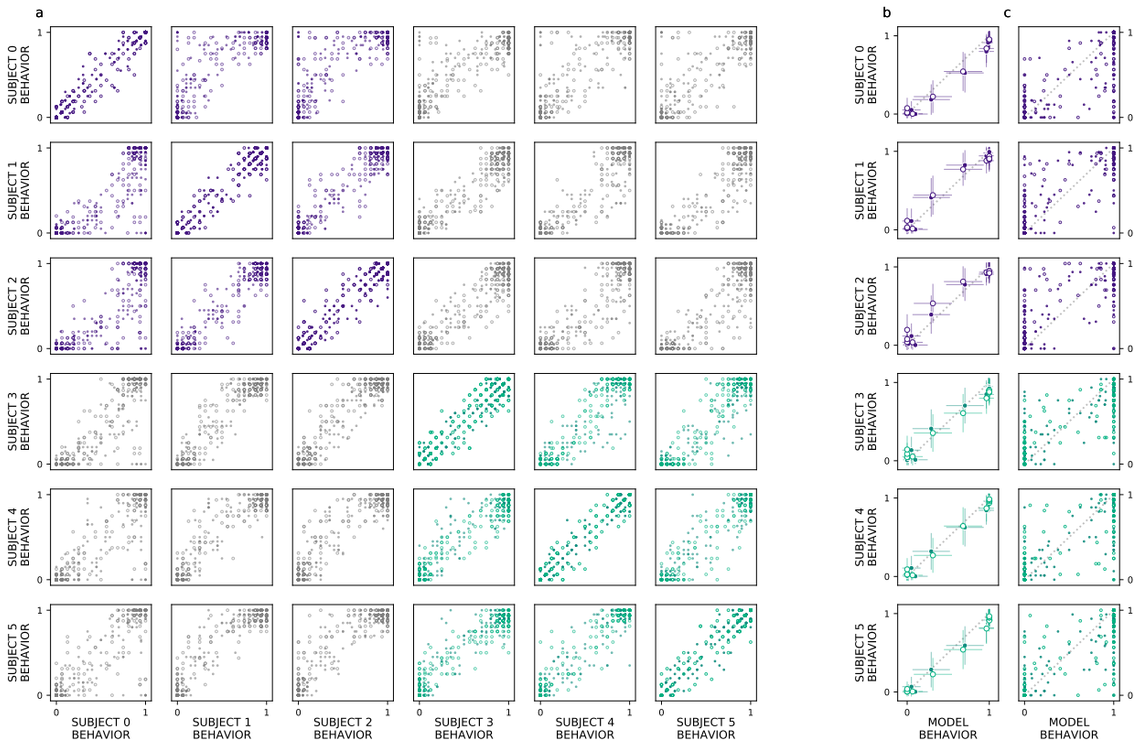

Figure 3: **VVS model fits subject behavior for aggregate but not image-level metrics. (a)** We estimate within- and between-subject reliability for both PRC-intact (purple, n=3) and -lesioned (green, n=3) choice behavior, across images and morph levels (small and large circles, respectively), experiment one (closed circles) and two (open circles). Within-subject reliability is visualized on the diagonal, off diagonal are the between-subject estimates for PRC-intact (purple), -lesioned (green), and between groups (grey). **(b)** Across morph levels, the VVS model predicts subject choice behavior across experiments. **(c)** The VVS model does not, however, predict subject choice behavior at the level of individual images.

when aggregated across morph levels, 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}) subjects (Fig. 3b). Interestingly, The model's fit to subject behavior is indistinguishable from the distribution of between-subject reliability estimates (Fig. 3b; median of the empirical $\text{P}(\text{model}|\text{reliability}_{\text{between-subject}}) = .592$) suggesting that the model exhibits subject-like behaviors at this resolution (between-subject reliability distributions visualized in Supplemental Fig. S2a). While our modeling approach is also able to predict image-level choice behaviors—again, for both PRC-lesioned (e.g. subject 3: $R^2 = 0.86$ $\beta = 0.81$, $F(1, 438) = 52.79, P = 5$ x 10^{-192}) and -intact subjects (e.g. subject 1: $R^2 = 0.87$ $\beta = 0.88$, $F(1, 438) = 53.24, P = 2$ x 10^{-193})—model behavior is significantly unlikely to be observed under the distribution of between-subject reliability estimates (Fig. 3c; median of the empirical $\text{P}(\text{model}|\text{reliability}_{\text{between-subject}}) = 0$). That is, while statistically significant, the model does not exhibit 'subject-like' image-level choice behaviors (between-subject reliability distributions visualized in Supplemental Fig. S2b).

3 Discussion

Theories surrounding neural function are commonly evaluated using informal, descriptive accounts of experimental properties. In the case of PRC involvement in visual object perception. terms like 'feature ambiguity' and 'stimulus complexity' have served as proxies for perceptual demands that can not be met by canonical visual cortices alone. Stimuli in Eldridge et al., 2018 were designed in accordance with these descriptive terms, owing to well-documented deficits observed in the PRC-lesion literature (Lee et al., 2005; Murray et al., 2007). Given the observation that PRC-lesioned subjects are not impaired on this ostensibly PRC-dependent perceptual task, the original authors reasonably conclude that PRC is not involved in perception. However, using more formal methods has lead us to a different interpretation of these same data: While these experiments can be *described* in terms that evoke PRC dependence, our results suggest that a linear readout of the VVS should be sufficient perform these stimulus sets. More concretely, we observe a striking correspondence between a computational proxy for the VVS and subject choice behavior—both PRC-intact and -lesioned groups. These results suggest that these experiments are in fact not diagnostic of PRC dependence, as perceptual processing beyond the ventral visual stream should

not be necessary to achieve neurotypical performance.

We emphasize that this work can not be interpreted as evidence *supporting* PRC involvement in perception. The available data are consistent with both perceptual-mnemonic and strictly mnemonic accounts of PRC function. Instead, we suggests that apparent inconsistencies in the experimental literature may be due to reliance on informal, descriptive accounts of perceptual demands. It may come as no surprise that these descriptive accounts do not reliably characterize neural function; indeed, the claim that cognitive constructs do not map cleanly onto the neuroanatomical divisions within the mammalian brain is central to the perceptual-mnemonic hypothesis (Murray and Wise, 2012). Instead of simply showcasing the limitations of informal, descriptive accounts of experimental variables, we have provided an extensible alternative: biologically plausible, stimulus-computable frameworks that predict animal behavior directly from experimental stimuli. Our computational proxy for the VVS serves, in effect, as a null model for PRC involvement in perception. This enables us to evaluate competing theories of neural function in concrete, falsifiable terms. Still, significant gaps between these computational approaches and animal behavior remain, as our image-level analyses makes clear. As such, understanding the resolution at which these methods approximate subject behaviors is critical for any modeling efforts. Nonetheless, we believe that this approach has wide applicability for better understanding the diverse behaviors supported by the medial temporal lobe.

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.

Bonnen, T., Yamins, D. L., & Wagner, A. D. (2020). When the ventral visual stream is not enough: A deep learning account of medial temporal lobe involvement in perception. *Available at SSRN 3758206*.

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. (2002). The organization of visual object representations: A connectionist model of effects of lesions in perirhinal cortex. *European Journal of Neuroscience*, *15*(2), 355–364.

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.

Bussey, T. J., Saksida, L. M., & Murray, E. A. (2006). Perirhinal cortex and feature-ambiguous discriminations. *Learning & Memory*, *13*(2), 103–105.

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.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

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.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

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., & Baxter, M. G. (2006). Cognitive neuroscience and nonhuman primates: Lesion studies. *Methods in mind*, *43*, 69.

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.

Murray, E. A., & Richmond, B. J. (2001). Role of perirhinal cortex in object perception, memory, and associations. *Current opinion in neurobiology*, *11*(2), 188–193.

Murray, E. A., & Wise, S. P. (2012). Why is there a special issue on perirhinal cortex in a journal called hippocampus? the perirhinal cortex in historical perspective. *Hippocampus*, *22*(10), 1941–1951.

Norman, G., & Eacott, M. (2004). Impaired object recognition with increasing levels of feature ambiguity in rats with perirhinal cortex lesions. *Behavioural brain research*, *148*(1-2), 79–91.

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.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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.

Suzuki, W. A., & Baxter, M. G. (2009). Memory, perception, and the medial temporal lobe: A synthesis of opinions. *Neuron*, *61*(5), 678–679.

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.

4 Methods

4.1 Validating model fit to electrophysiological responses

We begin with a task-optimized convolutional neural network, pre-trained to perform object classification. We use one instance of a task-optimized convolutional neural network (VGG16 (Simonyan and Zisserman, 2014), implemented in pytorch and pre-trained to perform object classification on a large-scale object classification dataset (Deng et al., 2009), but note that these results are reliable across model instances—with our own analysis on distinct models (e.g. resnets; He et al., 2016) reiterating what has been demonstrated in the literature (Rajalingham et al., 2018; Schrimpf et al., 2020). To identify a model layer that best fits IT cortex, we utilize previously collected (Majaj et al., 2015 electrophysiological responses from macaque V4 and IT cortex, along with the stimuli that elicited these responses. Using 'medium' and 'high' variation images from this data set, we convert each image from greyscale to RGB then resize it to accommodate model input dimensions (224x224x3). We pass each image to the model and extract responses from all layers (e.g. convolutional, pooling, and fully connected layers), vectorize each layer's output. We randomly segment

these model responses to each image into training and testing data using a 3/4th split. Thus, we use multi-electrode responses from macaque V4 and IT to a set of image, and model responses to those same images. For each layer, we learn a linear mapping between vectorized model responses and a single electrode's responses to the training images, using sklearn's implementation of PLS regression (with five components).

We evaluate this mapping between model and neural responses by computing the Pearson's correlation between model-predicted responses and observed responses for each electrode across all test images. For each layer, this results in a single correlation value for each electrode, which we repeat over all electrodes. This results in a distribution corresponding to that layer's cross-validated fits to population-level neural responses, both for electrodes in IT and V4. We compute the split half reliability for V4 ($r = .63 \pm .22$STD) and IT ($r = .73 \pm .24$STD) across neurons in each region. We then divide the distribution of cross-validated fits to IT and V4 by the reliability in each region—as a noise-corrected adjustment. This results in a single score—the noise-corrected, median cross-validated fit to both IT and V4—which we repeat across all layers. We also determine each layer's differential fit with primate IT, Δ_{IT-V4}, by taking the difference between the model's fit to IT and V4. Early model layers (i.e. first half of model layers) better predict neural responses in early (V4) regions of the visual system ($t(8) = 2.70, P = .015$), with peak V4 fits occuring in pool3 (noise-corrected $r = .95 \pm .30$STD) while later layers (e.g. second half of model layers) better predict neural responses in more anterior (IT) regions ($t(8) = 3.70, P = .002$), with peak IT fits occuring in con5_1 (noise-corrected $r = .88 \pm .16$STD); differential fit to IT cortex increases in with model depth (Δ_{IT-V4}; $\beta = .98$, $F(1, 17) = 20.91$, $P = 10^{-13}$).

4.1.1 Correspondence between model and VVS-supported performance

Here we directly compere model performance with VVS-supported performance: Instead of fitting model responses directly to electrophysiological recordings in high-level visual cortex, as above, we evaluate the similarity between the performance supported by the model and high-level visual cortex. For this comparision we again use electrophysiological responses previously collected from macaque IT cortex (Majaj et al., 2015. We independently estimate model and VVS-supported performance on stimulus set composed of concurrent visual discrimination trials, using a modified leave-one-out cross validation strategy. We then determine the model-VVS fit over the performance estimates, as developed in Bonnen et al., 2020. Each concurrent visual discrimination trial is composed of three images: two images contain the same object$_i$, randomly rotated and projected onto an artificial background; the other image (the 'oddity') contains a second object$_j$, again presented at a random orientation on an artificial background. For each trial, the task is to identify the oddity—that is, the object which does not have a pair—ignoring the viewpoint variation across images (The experiment can be demo'd online at https://stanfordmemorylab.com:8881/high-throughput_data_collection/index.html).

We use a modified leave-one-out cross validation strategy to estimate model performance across stimuli in this experiment. For a given sample$_{ij}$ trial we construct a random combination of three-way oddity tasks to be used as training data; we sample without replacement from the pool of all images of object$_i$ and object$_j$, excluding only those three stimuli that were present in sample$_{ij}$. This yields 'pseudo oddity experiments' where each trial contains two typical objects and one oddity that have the same identity as the objects in sample$_{ij}$ and are randomly configured (different viewpoints, different backgrounds, different orders). These 'pseudo oddity experiments' are used as training data. We reshape all images, present them to the model independently, and extract model responses from an 'IT-like' model layer (in this case, we use fc6 which has a similar fit to IT as conv5_1 but fewer parameters to fit in subsequent steps). From these model responses, we train an L2 regularized linear classifier to identify the oddity across all ($N = 52$) trials in this permutation of pseudo oddity experiments generated for sample$_{ij}$. After learning this weighted, linear readout, we evaluate the classifier on the model responses to sample$_{ij}$. This results in a prediction which is binarized into a single outcome $\{0 \mid 1\}$, either correct or incorrect. We repeat this protocol across 100 random sample$_{ij}$s, and average across them, resulting in a single estimate of model performance for each pair$_{ij}$

To relate model performance with the electrophysiological data, we repeat the leave-one-out cross-validation strategy described above, but in place of the fc6 model representations we run the same protocol on the population-level neural responses from IT and V4 cortex. We perform all analyses comparing model and VVS-supported performance at the object level: for each object$_i$ we average the performance on this object across all oddities (i.e. object$_j$, object$_k$, ...) resulting in a single estimate of performance on this item across all oddity tasks ($N = 32$). We can compare model performance with both VVS-supported performance and PRC-intact (human) performance on these same stimuli, using data from Bonnen et al., 2020. On this dataset, PRC-intact human behavior outperforms a linear readout of macaque IT (Fig 1c: $\beta = .24$, $t(31) = 9.50$, $P = 1\text{x}10^{-10}$), while IT significantly outperforms V4 ($\beta = .18$, $t(31) = 6.56$, $P = 2\text{x}10^{-7}$). A computational proxy for IT

demonstrates the same pattern, predicting IT-supported performance ($\beta = .81$, $F(1, 30) = 13.33$, $P = 4\text{x}10^{-14}$), outperforming V4 ($\beta = .26$, $t(31) = 8.02$, $P = 5\text{x}10^{-9}$), and being outperformed by PRC-intact participants ($\beta = .16$, $t(31) = 5.38$, $P = 7\text{x}10^{-6}$).

Next we evaluate the relationship that both the model and IT-supported performance have with human reaction time. While PRC-intact subjects outperform IT, we nonetheless expect that IT-supported performance will relate to PRC-intact accuracy: For trials where IT-supported performance is greater, there is less of a demand for extra-IT processing, and more accurate PRC-intact accuracy. In line with this expectation, there is a reliable relationship between IT-supported and PRC-intact performance ($\beta = .26$, $F(1, 30) = 8.49$, $P = 2\text{x}10^{-9}$). Additionally, for each item, the difference between IT-supported and PRC-intact performance is predicted by reaction time ($\beta = .81$, $F(1, 31) = 7.44$, $P = 3\text{x}10^{-8}$). That is, PRC-intact human participants require more time to choose among items that are not linearly separable in IT, in a way that scales inversely with IT-supported performance. Critically, this relationship is observed for model performance ($\beta = .72$, $F(1, 31) = 5.62$, $P = 4\text{x}10^{-6}$) but not V4-supported performance ($\beta = -.08$, $F(1, 31) = -0.41$, $P = .682$). Taken together, these results suggest that our model is able to approximate VVS-supported performance. This model exhibit the same pattern of performance as IT, in relation to V4 and PRC-intact human behaviors—not only with respect to accuracy, but the reaction time inherent in outperforming the VVS.

4.2 Consistency estimates

We estimate within- and between-subject consistency using a common protocol. For the given resolution of analysis (either morph- or image-level), we require multiple presentations of the same items. For the morph-level analysis, which aggregates stimuli within 'morph levels' (e.g. all stimuli that are designed to be 0% dog, 10%, etc.), all stimulus sets meet this criterion. There are, however, multiple experiments that do not contain sufficient data to perform the image-level analysis, which requires multiple presentations of each stimulus; experiment four contains only one presentation of each stimulus, precluding it from our consistency analyses, and experiment three contains only 4 repetitions, which is insufficient for reliable within- and between-subject consistency estimates. Thus, we restrict our consistency estimates to experiments one (10 repetitions per image) and two (8 repetitions per image). All repetition counts are evident in Fig. 2.

We estimate all consistency metrics over 100 random split-halves iterations. For each iteration, across all items within a given resolution (where items can refer to either a give morph percent, for the morph-level analysis, or a given image, for the image-level analysis), we randomly split choice behavior into two random splits. In the image-level analysis, for example, for each image x_i within the set of n images, we randomly select half of all repetitions of x_{i_1}, and compute the mean of this random sample $\bar{x}_{i_1}$, for all n images. We repeat this procedure for the remaining half, $\bar{x}_{i_2}$, across all n images. Thus, we have two n dimensional vectors, $\vec{v}_1$ and $\vec{v}_2$, randomly sampled from the full distribution of choice behaviors. We use R^2 as a measure of fit between each split half. We repeat this measure over 100 iterations, resulting in a distribution of split-half fits. For within-subject consistency metrics, we generate each split by randomly sampling from a single subjects' choice behaviors. For the between subject consistency metrics, we compute $\vec{v}_1$ from subject$_i$s choice behavior, and $\vec{v}_2$ from subject$_j$s choice behavior, using the same protocol described above.

5 Supplemental Information

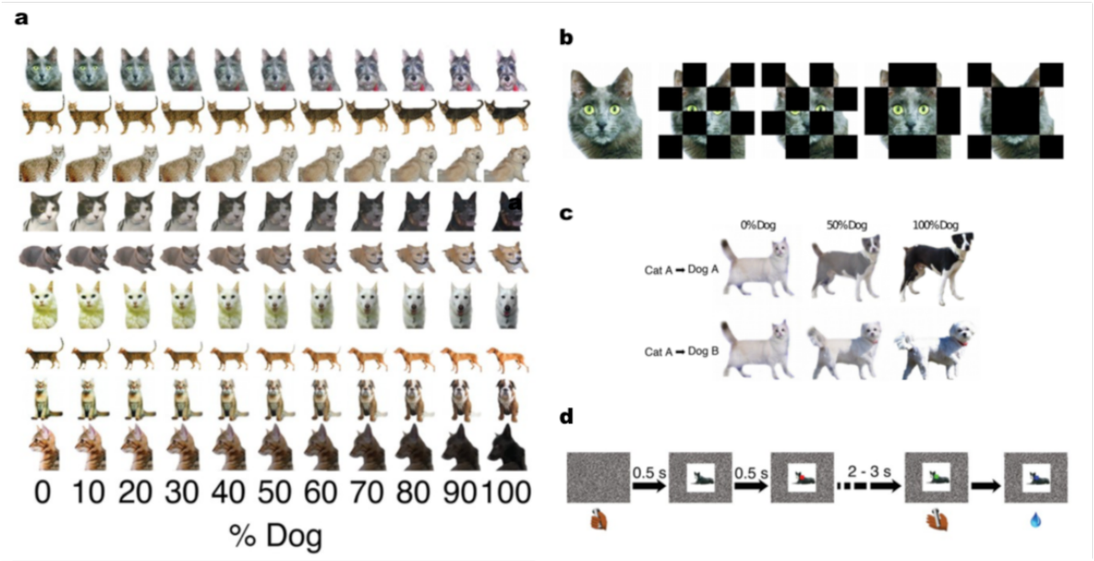

Figure S1: **Experimental stimuli and protocol from Eldridge et al., 2018. (a)** Example stimuli from experiment one, illustrating multiple instances of stimuli across morph levels. **(b)** Example stimuli used used for masked morphs, in experiment 3. **(c)** Example stimuli used for 'crossed morphs' in experiment 2. **(d)** Protocol for all experiments. Subject's initiate each trial with a lever press. A stimulus is presented, followed by a red dot at the central field of view. Subjects could avoid an extended inter-trial delay by releasing the bar in the first interval (signaled by a red target) for stimuli that were less than 50% dog, and were rewarded for releasing the bar in the second interval (signaled by a green target) for stimuli that were more than 50% dog. This amounts to an asymmetrical reward structure. They were rewarded randomly for releasing during the green interval for 50 – 50 morphs.

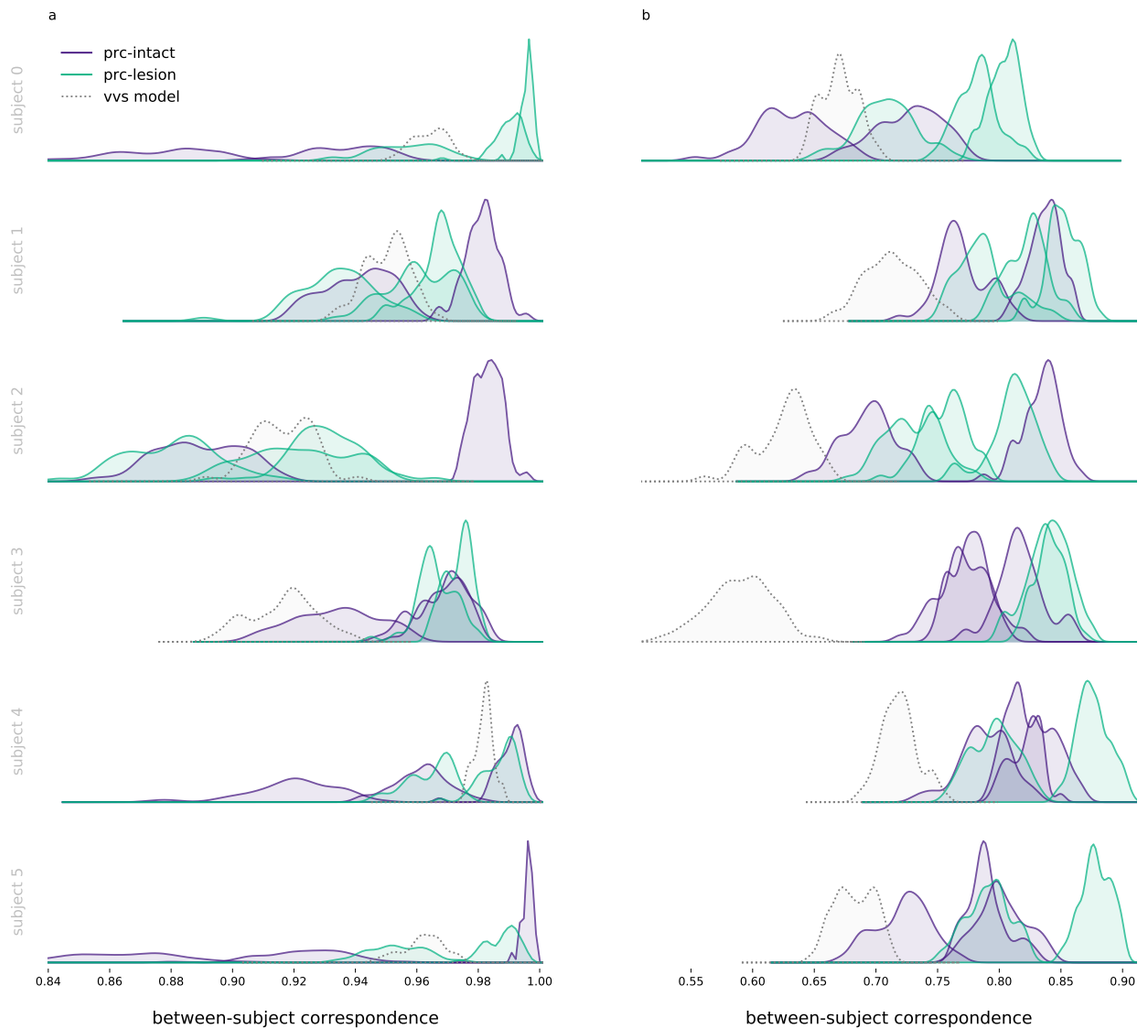

Figure S2: **Between-subject correspondence for aggregate and image-level analyses.** We estimate the correspondence between subject choice behaviors over 100 split-half iterations, using R^2 as a measure of fit. Each row contains a given subjects' (e.g. subject 0, top row) correspondence with all other subject choice behaviors, for PRC-intact (purple) and -lesioned (green) subject behaviors. We perform this analysis at two resolutions: **(a)** averaging performance across all images within each morph level (e.g. 10%, 20%, etc.) used in Fig. 2, and **(b)** the image-level analysis used in Fig. 3a. For each subject, we also estimate the correspondence with a computational proxy for the VVS over 100 iterations, again using R^2 as a measure of fit at each iteration. The resulting distribution of subject-model correspondence across iterations is visualized in grey (dashed). For the morph-level analysis, the model choice behavior is 'subject-like'; the distribution of model-subject correspondence is within the distribution of between-subject correspondence. However, at the resolution of single images, model choice behavior is not subject-like; the observed model fit to each subject is not likely observed under the between subject reliability distributions.