Model manifold analysis suggests the human visual brain is less like an optimal classifier and more like a feature bank

Colin Conwell
MIT CSAIL
conwell@mit.edu

Michael F. Bonner
Johns Hopkins University
mfbonner@jhu.edu

Abstract

What do deep neural network (DNN) models actually tell us about the computational principles of visual information-processing in the biological brain? A now common finding in visual neuroscience is that many different kinds of DNN models – with different architectures, tasks, and training diets – are all comparably performant predictors of image-evoked brain activity in the ventral visual cortex. This relative parity of highly diverse models may at first seem to undermine the common intuition that we can use these models to infer the computational principles that govern the visual brain. In this work, we show to the contrary that comparable brain-predictivity does not preclude the differentiation of these same models in terms of the underlying manifold geometries that define them. To do this, we assess 12 manifold geometry metrics computed across a diverse set of 117 DNN models, curated to include multiple tasks, architectures, and input diets. We then use these metrics to predict how well each model aligns with occipitotemporal cortex (OTC) activity from the human fMRI Natural Scenes Dataset. We find that *manifold signal-to-noise ratio* (a metric previously associated with few-shot learning) is a robust predictor of downstream brain-alignment and supersedes both other manifold geometry metrics (i.e. *manifold capacity*) and downstream task-performance (e.g. top-k recognition accuracy) across multiple different image sets (e.g. ImageNet21K versus Places365) and controlled model comparisons (e.g. assessments across ImageNet-1K trained architectural variants only). These results add to a growing body of evidence that the ventral visual stream serves as a basis set (or feature vocabulary) for object recognition rather than as the actual locus of recognition *per se*.

1 Background

High-level ventral visual cortex is widely considered the primary substrate of object recognition in the human brain [1; 2; 3]. This notion has been further reinforced in recent years by the now-seminal finding that deep neural network (DNN) models supporting image classification are the most predictive models of ventral visual brain activity to date [4; 5; 6]. One lingering issue with this formulation, however, is the under-specification of what we mean by the word "recognition" as implemented by the biological brain. What *precise* neural-computational mechanisms instantiate this process? And how might we use deep neural network models to elucidate them? In models, the computational mechanisms of 'recognition' are explicitly (i.e. algorithmically) defined. Canonical, 'end-to-end' image classification models (e.g. AlexNet [7]) 'recognize' images by the direct, nonlinear mapping of a tensorized image onto one-hot-encodings (point indices) of human-defined category labels. Self-supervised, contrastive-learning models (e.g. SimCLR [8]) and masked decoding models (e.g. DINO, [9]) also support recognition, but do so first by learning (without labels) the invariances and selectivities that define the axes of an input image space more generally. 'Zero-shot'- multimodal classifiers (e.g. CLIP [10]) 'recognize' images by way of a top-k similarity score between the language embeddings of category labels and the vision embeddings of candidate images.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: DBM.

Identifying which of these (or many other) motifs is most brain-like has been a process somewhat confounded by the fact that many (if not most) of these models often achieve roughly comparable (often high) measures of ‘representational alignment’ [11] to human visual cortex activity recorded via neuroimaging and neurophysiology alike [6; 12; 13; 14; 15]. One response to this relative parity, inspired by the neuro-connectionist research programme [16], has been the use of controlled model comparisons designed to better parse the differences between models by *grouping them* in ways empirically isolate the representational influence of broader computational design principles such as sensory diet (training data), architecture, and task [17]. In this work, we attempt to better parse the differences between models more directly by *sorting them* along the scalar axes of representational structure defined by manifold geometry analysis [18; 19; 20]. Where the neuro-connectionist programme asks *what* (i.e. which design constraints predict brain-alignment), and representational similarity analysis asks *how much* (i.e. to what degree do model and brain representations covary), manifold geometry analysis asks *how*: specifically, how does the *structure* of a representation (its separability, compactness, dimensionality) relate to the downstream computations (e.g. classification, generalization) it supports? By linking structural properties to functional outcomes, manifold analysis provides a mechanistic lens that representational alignment metrics alone do not directly afford. In so doing, we attempt as well to more directly interrogate how this structure relates to the *function* of object recognition, and what our neural networks may or may not be telling us about the nature of this structure in the human ventral visual system.

Precedent + parallel work: The use of manifold geometry to understand representational structure (brain-like or otherwise) is a technique steeped in the history of computational modeling and theoretical neuroscience [see e.g. 21], as well as an actively developing area of interest. Recent work in this domain has seen manifold geometry analysis applied to *speech* recognition in (biological and artificial) auditory neural networks [22]; in theoretical work on the limits or underpinnings of representational alignment more generally [23; 24]; to the characterizing of task-optimized models predicting activity in macaque visual cortex [25]; and (most directly relevant to the current analysis) to the characterizing of representational structure in human ventral visual system activity sourced from the Natural Scenes Dataset [26; 27]. St-Yves et al. [26] in particular is notable for the use of the same manifold signal-to-noise and dimensionality metrics [20] we leverage in this analysis, deployed in service of characterizing how geometry varies across brain region. Our work builds and extends on these approaches by explicitly yoking the more distal, ecological insights of the neuro-connectionist, many-model / controlled comparisons approach with the more proximate, structural insights of manifold geometry - and in so doing, to get closer to unifying the goals of ‘prediction’ and ‘understanding’ that so often seem in tension in the application of DNN models to neuroscience data [28].

2 Methods

Model & metric curation: Our general approach was to first curate N=12 metrics of manifold geometry from two distinct lines of work: N=4 metrics from Chung et al. [18] (derived from replica mean field theory) and N=8 metrics from Sorscher et al. [20] (derived from principal components analysis). We then curated a set of N=117 candidate deep neural network models spanning different visual diets, architectures, and tasks, based on the work of Conwell et al. [17].

Measuring brain-alignment: Model-to-brain predictivity scores (henceforth, brain-alignment) was computed for each model first by computing the average voxelwise encoding score over a 50% train split (N=500) of the ‘Shared NSD1000’ image set [29] for each individual layer with field-standard regularized ridge regression. Models were then compared on the basis of the average encoding scores over the test split (N=500) images for the most brain-like layer (as determined via nested cross-validation on the train split).

Manifold geometry analysis: Manifold geometry metrics (from both Chung et al. [18] and Sorscher et al. [20]) were computed using N=50 ‘concept manifolds’ (the representational matrices of 50 images from each of 50 category labels from ImageNet-1K/21K [30; 31], or Places365 [32]). This 50×50 sampling scheme (2,500 total images) follows the protocol established in Stephenson et al. [22] and provides sufficient statistical power to estimate manifold properties reliably; prior work has shown these estimates to be stable under bootstrap resampling of both exemplars and categories [18; 20]. To ensure proper evaluation of these metrics on our candidate representations, we used the exact GitHub source code associated with the defining works of each: schung039/neural-manifolds-replicaMFT [18; 22] and bsorsch/geometry-fewshot-learning [20].

Manifold metric descriptions: The N=4 metrics from Chung et al. [18] characterize manifold geometry via replica mean field theory: *capacity* measures the maximum number of linearly separable dichotomies a linear classifier can implement; *radius* measures the average manifold extent relative to centroid separation; *dimensionality* measures the effective number of dimensions each manifold occupies; and *correlation* captures the average pairwise correlation between manifold centroids. The N=8 metrics from Sorscher et al. [20] are derived from PCA: *signal-to-noise ratio* (SNR) quantifies between-class variance relative to within-class variance in the principal subspace, directly predicting few-shot learning performance; *signal* and *bias* decompose SNR into its numerator and denominator components; *effective/between-concept/within-concept dimensionality* capture the intrinsic dimensionality of the full space, between-class subspace, and within-class subspaces respectively; *within-concept radius* measures average manifold compactness; and *SNR overlap* measures the geometric overlap between signal and noise subspaces.

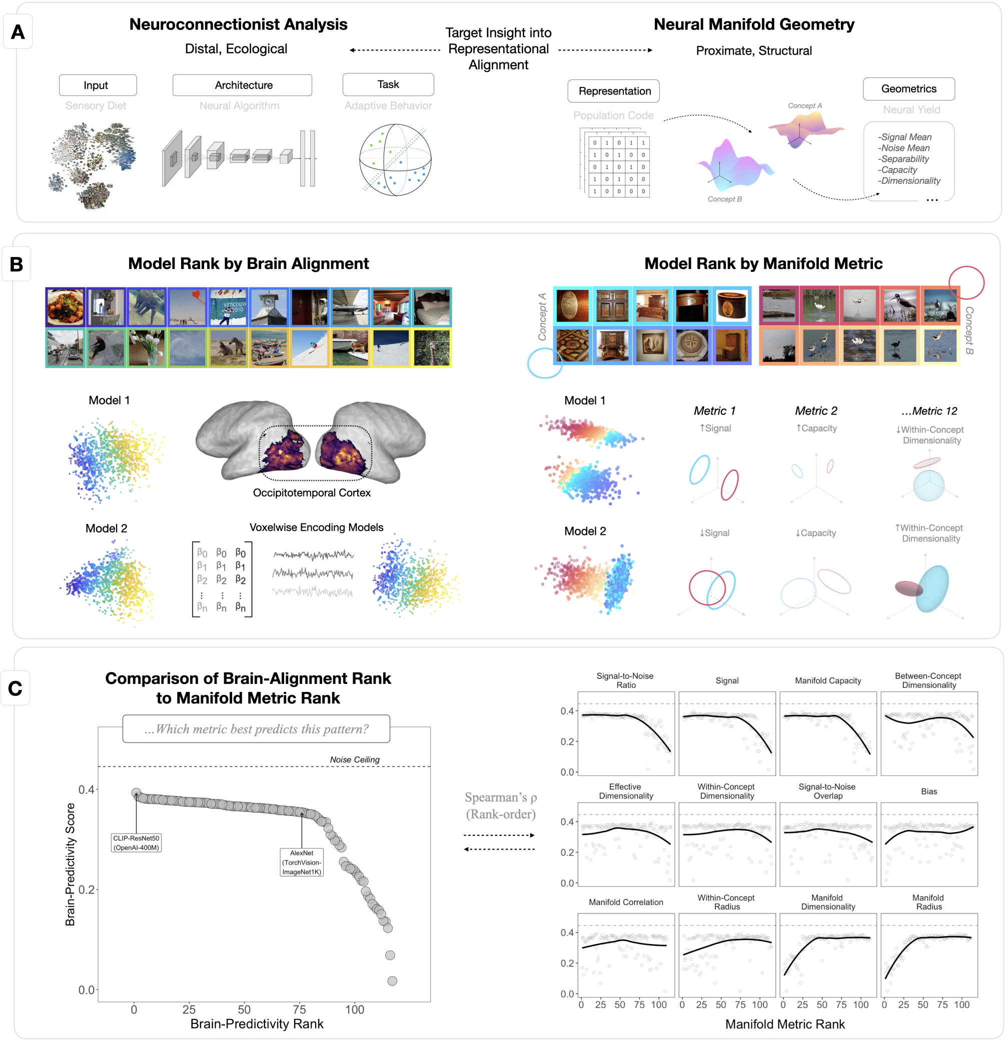

Figure 1: An overview of our motivation, methodology, and primary analysis. An extend caption with more details on each panel is available in the paragraph below.

.

Overview (Figure 1) extended caption: In **A**, we schematize the key factors that define and contrast the two animating frameworks of our model-to-brain comparisons. With its focus on the influence of input, architecture, and task, neuroconnectionist-style analysis [16; 17] seeks to use models as proxies of the design constraints (i.e. 'pressures') that could in principle have shaped the emergence of the representational structure we observe in the biological brain. Neural manifold geometry [19; 20] seeks to use models as a direct empirical substrate for probing how the differences in the structure of representation (both within and across models) contribute to differences in downstream behavior (i.e. classification, or in this case, brain-alignment / predictivity). In **B**, we schematize our primary method for using manifold geometry to interpret brain-alignment in high-level ventral visual stream (occipitotemporal cortex, or OTC). We first rank models in terms of their brain-alignment scores by computing average voxelwise encoding scores with field-standard cross-validated ridge-regressions for the 'Shared-NSD1000' [29]. We then take these same models and rank them according to each of our curated manifold geometry metrics computed over 'concept manifolds' (representational matrices) composed from ImageNet categories. In **C**, we show the primary outcome of this analysis: A rank-order comparison (Spearman's ρ correlation) between the rank of each model according to its brain-predictivity, and the rank of each model according to its associated manifold geometry metrics. As shown, the trend in the brain-predictivity plot on the left (with brain-predictivity score in units of Pearson's r on the y axis, and brain-predictivity rank on the x axis) is better and worse captured by the various metrics in the subplots on the right (with manifold metric rank in place of the brain-predictivity on the x axis), which are sorted from top to bottom by Spearman's ρ.

3 Results + Analysis

Primary analysis: Explaining Brain-Alignment with Manifold Geometry: How well do differences in underlying manifold geometries explain downstream brain-alignment? To answer this, we compute the rank-order correlation ($\rho_{Spearman}$) between the models' brain-predictivity (mean encoding) scores and each of our 12 manifold metrics (e.g. manifold capacity, effective dimensionality) computed from the N=50 concept manifolds of ImageNet-21K used in [22]; results from this analysis are shown in Figure 1C and Table 1.

Manifold Metric	$r_{\text{Spearman}} \pm 95\% \, CI$	p
Chung et al.		
Capacity	**0.643 [0.527, 0.796]**	**<.001**
Correlation	-0.077 [-0.266, 0.107]	0.42
Dimensionality	-0.594 [-0.747, -0.471]	<.001
Radius	-0.721 [-0.856, -0.62]	<.001
Sorscher et al.		
Signal-to-Noise Ratio	**0.774 [0.706, 0.873]**	**<.001**
Signal	0.569 [0.445, 0.724]	<.001
Between-Concept Dimensionality	0.269 [0.103, 0.45]	0.003
Effective Dimensionality	0.154 [-0.018, 0.336]	0.097
Signal-to-Noise Overlap	0.133 [-0.04, 0.317]	0.153
Within-Concept Dimensionality	0.073 [-0.104, 0.26]	0.433
Within-Concept Radius	-0.239 [-0.429, -0.068]	0.009
Bias	-0.242 [-0.409, -0.088]	0.009

Table 1: Rank-order correlations between brain-predictivity and manifold geometry metrics, with bootstrapped 95% confidence intervals shown in brackets; results from the Primary Analysis. Manifold capacity and SNR (the main metrics we probe in our follow-up experiments) are in **bold**.

6 / 12 metrics showed significant rank-order correlation with brain-predictivity at $p < 0.01$. 3 of these metrics (manifold signal-to-noise ratio, capacity, and signal) showed significant, positive correlations. 3 (manifold radius, dimensionality, and within-concept radius) showed significant, negative correlations. The overall strongest predictor of downstream brain-predictivity was the manifold signal-to-noise ratio from Sorscher et al. [20], with a strikingly high rank-order correlation of ρ_{Spearman} [$\pm 95\% BCI$] = 0.798 [0.731, 0.895], p = 4.70e-27. Not far behind this, however, was the negatively correlating manifold radius metric from Chung et al. [18], at ρ = -0.724 [-0.857, -0.623], and the positively correlating manifold capacity metric from Chung et al. [18], at ρ = 0.616 [0.493, 0.779]. In sum, the primary takeaway from this first analysis might simply be that manifold metrics

can indeed be the strong predictors of brain-alignment that theorists and empiricist alike have long proposed they might be [20; 33; 23; 24; 34].

Querying manifold metric robustness & interpretability across experimental subconditions: The strong, significant correlation of *multiple* manifold metrics with downstream brain-alignment does, however, raise new questions. One of these is: When metrics describe otherwise divergent properties of a manifold's geometry (but both explain downstream brain-alignment to similar degrees), which geometry should we take to be more brain-like? In our survey, two metrics in particular – manifold capacity and manifold signal-to-noise ratio (henceforth SNR) – instantiate a rather palpable case of this algorithmic ambiguity. In the extreme (as in the final output layer of an end-to-end trained category-supervised DNNs), manifold capacity heralds the complete collapse of all category information into fully separate, single point-estimates of category identity Chung et al. [18]; Stephenson et al. [22]; Chung and Abbott [19] – in other words, an optimal classifier. In a similar extreme, manifold signal-to-noise ratio also collapses to perfect point-estimates of category identity (all signal, no noise). As described by Sorscher et al. [20], however, mid-to-high range values of signal-to-noise ratio better describe the conditions of better few-shot learning algorithms (something end-to-end category-supervised neural classifiers tend notably *not* to be). Which of these metrics, then, better describes the object-recognition-supporting representational motifs instantiated in our DNN models of the ventral visual stream? To better resolve this ambiguity, we assessed the correlations of manifold capacity and SNR to downstream brain-predictivity in a series of experimental sub-conditions designed both to test the robustness of these correlations and to disambiguate the somewhat competing representational hypotheses they entail. (Details from Experiments 1, 2 and 3 are shown in Tables 2 and 3.)

Experiment 1: Measuring robustness across model (sub)sets: In our first experiment, we assessed the correlations of manifold capacity and SNR in increasingly smaller, targeted subsets of our otherwise diversely sampled model set. The first subset we assessed was a subset of models we call the "high-performing" set: effectively all models above a notable visual elbow in brain-predictivity first described in [17], but seen also in our sample (see Figure 1C-Left). Quantifying this elbow with a segmented regression analysis yielding a breakpoint of ψ = 84.1 [72.4, 94.9], we defined this 'high-performing' set as the N=84 models with average encoding scores of r_{Pearson} = 0.336 or higher. Here, already, manifold capacity begins to diverge from manifold SNR in its rank-order correlation with brain-predictivity, with capacity diminishing to the point of non-significance at ρ = 0.14, p=0.222 and manifold SNR remaining substantial and significant at ρ = 0.50, p < .001. In an even smaller subset of models varying only in architecture (the N=53 category-supervised ImageNet-1K-trained models from the Torchvision model zoo [35]), the trend is similar, with ρ for manifold capacity diminishing to 0.16, p = 0.262 and manifold SNR remaining high at 0.471, $p < 0.001$. In short, manifold SNR persists as a predictor of downstream brain-alignment even in very restricted ranges; manifold capacity does not.

	Manifold Capacity	Signal-to-Noise Ratio
	ρ_{Spearman}, p	ρ_{Spearman}, p
All Surveyed Models	0.62, $p < 0.001$	0.80, $p < 0.001$
High-Performing	0.14, $p = 0.222$	0.52, $p < 0.001$
ImageNet1K-Supervised	0.16, $p = 0.262$	0.47, $p < 0.001$

Table 2: A comparison of the rank-order correlations between the manifold capacity and manifold signal-to-noise ratio (SNR) metrics across progressively smaller subsets of models; results from Experiment 1.

Experiment 2: Layer-wise analysis of category-recognition models: To better understand the difference between manifold capacity and manifold SNR we were observing in this smaller subset, we next probed variation in the correlation of manifold metric and brain-predictivity in layers beyond the most brain-predictive layer selected by our initial cross-validation, and in particular, the 'last hidden layer' feeding into the one-hot, category-encoding output. In effect, in this smaller subset of category-supervised models, this layer instantiates the representation most directly responsible for the 'recognition' behavior the model will output for any given input, and by association, is the layer we might presumably observe the highest covariance between manifold capacity and manifold SNR. And indeed, what we observe here is that the correlations of both manifold capacity and manifold

SNR with downstream brain-predictivity change dramatically. Here, in this final hidden vector of models trained to collapse category information onto the single points of the output layer, manifold capacity and manifold SNR are shown to be strong, significant, *negative* predictors of downstream brain-alignment, with ρ = -0.606, p < .001 and -0.663, respectively. This sign-reversal corresponds to substantial increases in the scalar values of both metrics relative to the most brain-predictive layers (with shifts in the max values of manifold capacity increasing 182.17% from .129 to .235 and signal-to-noise ratio increasing 209.12% from 4.131 to 8.639.) What is happening, in effect, is that the more the models are successful in collapsing category information to single point-estimates at this final hidden layer, the less predictive of downstream brain-alignment they will be. Notably, the trend is similar if as with our manifold metrics, we correlate the ImageNet-1K classification accuracy of these models with downstream brain-predictivity, a trend we find (in line with recent work, e.g. [36]) to be strongly negative across our 53 models (ρ = -0.63, p < .001).

Experiment 3: Differential, IID/OOD concept manifold sampling: In a final experiment, we recomputed each of our manifold metrics with two new sets of N=50 concept manifolds (N=50 test set images each): one from the object categories of ImageNet1K (versus the ImageNet21K sample we use in our main analysis, following the protocol and codebase of Stephenson et al. [22]), and another from the *scene* categories of Places365. The logic here is that these instantiate two different levels of 'generalization' for our category-supervised ImageNet-1K-trained models, one nearer (in-distribution, IID), one farther (*possibly* out-of-distribution, OOD). Manifold capacity, in this regime, should be higher for those concepts that are IID (i.e. the ImageNet1K sample) than OOD (the Places365), again instantiating the progressive tightening of category information towards single point-estimates. Manifold SNR, on the other hand, will also decrease. But supporting few-shot-learning, as it nominally does, manifold SNR will also maintain information that bridges the gaps between separable categories. Is this the kind of information that boosts manifold SNR's explanatory power for downstream brain-alignment? The results of this experiment suggest it might be: In the most brain-predictive layers, for example, we observe that manifold SNR remains a significant, positive predictor of downstream brain-alignment in both the new ImageNet1K concept manifold sample *and* in the Places365 sample, with ρ = 0.457 p < .001 and 0.319, p = .002, respectively. Manifold capacity (as in the ImageNet21K sample) is *not* significantly predictive of downstream brain-alignment in either of these cases.

	Manifold Capacity		Signal-to-Noise Ratio	
	Range	$r_{\text{Pearson}},\ p$	Range	$r_{\text{Pearson}},\ p$
ImageNet1K				
Best Layer	0.049 - 0.129	-0.070, $p = 0.637$	2.038 - 4.131	0.457, $p < 0.001$
Last Layer	0.098 - 0.235	-0.606, $p < 0.001$	1.622 - 8.639	-0.663, $p < 0.001$
ImageNet21K				
Best Layer	0.052 - 0.125	-0.132, $p = 0.373$	2.092 - 4.062	0.468, $p < 0.001$
Last Layer	0.095 - 0.160	-0.515, $p < 0.001$	1.833 - 4.467	-0.294, $p = 0.032$
Places365 (Scenes)				
Best Layer	0.047 - 0.095	-0.145, $p = 0.327$	1.677 - 3.446	0.319, $p = 0.02$
Last Layer	0.070 - 0.108	-0.515, $p < 0.001$	1.305 - 2.956	-0.158, $p = 0.259$

Table 3: Comparisons of manifold capacity and signal-to-noise ratio between the peak (i.e. most brain-predictive) layer and the last layer in 3 different probe datasets. $r_{Pearson}$ is the correlation between the manifold metric value and brain-predictivity (mean encoding score) of each layer; results from Experiments 2 and 3.

4 Discussion

What factors make for a 'good' neural network model of the visual brain? Since the adoption in visual neuroscience of the task-optimized deep convolutional neural network model more than a decade ago [37], the dominant – and in some ways most empirically defensible answer – has largely been 'prediction': Better models of the visual brain are those models whose internal representations most accurately *predict* the activity patterns of the biological brain. For those seeking downstream control or causal perturbation of biological systems [38; 39], this answer may be sufficient. For those seeking 'understanding', the search remains for other forms of explanatory variables that supplement raw prediction with the parsimony of theories articulable in formal or natural language [40; 16; 34].

In this work, we attempt to instrumentalize the emergent framework of neural manifold geometry [19] to better understand the underlying *structural* factors that make certain models of ventral visual cortex 'better' (i.e. more predictive of brain activity) than others – in effect, by using the metric scalars of manifold geometry to more directly link *representation* to *function*. Through this lens, we return to the seminal question of how the representation in ventral visual cortex relates to the function of object recognition (i.e. the 'readout' of a category label). First validating the second-order predictive power of manifold geometry metrics (i.e. the strength of their rank-order correlation with downstream brain-alignment), we find multiple candidate metrics that nevertheless instantiate divergent hypotheses about the object-recognition-supporting representations of the ventral stream. Testing these hypotheses in a series of experimental subconditions, we find that a metric (manifold capacity) whose value scales with representational convergence toward separable, single point-estimates of category identity is less robust in predicting downstream brain-alignment than a metric (manifold SNR) that accounts for more graded forms of invariance and separability that still subserve recognition, but equally subserve the few-shot learning of new categories.

On the interpretation of OTC as "feature bank" versus "optimal classifier": We frame our findings in terms of this contrast not because we believe anyone expects occipitotemporal cortex (OTC) to literally function as an end-to-end classifier in the manner of a softmax output layer – indeed, even neurons in prefrontal regions associated with perceptual decision-making do not resemble classifier outputs [41]. Rather, the "optimal classifier" framing serves as a useful limiting case against which to interpret manifold geometry: if OTC representations were optimized primarily for categorical discriminability (maximizing capacity), we would expect a different pattern of results than if they were optimized for flexible, generalizable feature extraction (maximizing SNR) with variance that suggests a more limited tolerance of inter- and intra-class variation. Our results favor the latter interpretation, but we emphasize that this is a relative, not absolute, claim. The ventral stream likely serves multiple computational functions simultaneously, and our evidence does not preclude other interpretations (e.g., that OTC representations are optimized for some combination of recognition, scene understanding, action affordance, and memory encoding alike). What we can say with greater confidence is that the representational geometry of brain-predictive model layers diverges systematically from the geometry of layers optimized purely for classification.

Taken together, these results add to a growing body of evidence from across multiple experimental modalities suggesting that the ventral visual stream may be less like an optimal classifier (i.e. the locus of recognition itself) and more like a feature bank (i.e. the vocabulary of whatever compositional process the *active* act of recognition involves downstream). Whether it be the finding that the biological ventral stream in macaques may be more 'texture-like' than previously assumed on the basis of shape-based behavioral biases [42]; the greater impairment to object recognition from damage to medial temporal lobe than ventral stream in double dissociation neuropsychology experiments [43; 44]; the demonstration of emergent category-like topographic structure from self-organized maps learned over self-supervised natural image models [45]; or even more simply the now robustly reproduced finding (inherent to these results as well) that the later layers of category-recognizing deep neural network models are less brain-predictive than more intermediate layers [17], the ventral stream seems in many cases to be preserving information in ways that diverge from the strict motif of category invariance and separability that one might assume were one to assume the function of this region is the kind of classification subserved by the models that best predict its activity.

Limitations and future directions: Several limitations of the current work warrant acknowledgment. First, our operationalization of brain-alignment relies on encoding model predictivity (voxel-wise ridge regression); other alignment metrics such as representational similarity analysis (RSA), centered kernel alignment (CKA), or canonical correlation analysis (CCA) may yield different patterns of results [17, .c.f.], and investigating this is an important direction for future work. Second, the inferences we make about the manifold geometry of the brain are inferences we make by proxy of the brain-predictive models that are the actual targets of our analysis – future work should apply manifold geometry analysis directly to neural data where feasible [26]. Third, our model set, while diverse, does not include models trained on more ecologically valid datasets such as Ecoset [46], which better approximate the statistical structure of natural visual experience; including such models could strengthen claims about the ecological constraints shaping ventral stream representations.

Increasingly, however, we have seen that however much ‘proxy’ the inferences we make from models may be, these inferences are often directly convertible into the kinds of causal / perturbational empiricism that are the gold standard of scientific understanding for a target biological system. Concurrent work is already applying manifold geometry analysis and similar approaches towards the goal of robustifying brain-predictive models through explicit, manifold metric-guided representational alignment and neural control [27; 47]. Future work should continue to leverage the structural grip of manifold analysis with the high-throughput development of ever-more competent task-optimized models to accelerate the incisive style of this empiricism even further.

References

[1] James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007. doi: 10.1016/j.tics.2007.06.010. Publisher: Elsevier.

[2] J.J. DiCarlo, D. Zoccolan, and N.C. Rust. How Does the Brain Solve Visual Object Recognition? *Neuron*, 73(3):415–434, 2012. doi: 10.1016/j.neuron.2012.01.010. URL `https://doi.org/10.1016/j.neuron.2012.01.010`.

[3] Nancy Kanwisher. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, 107(25):11163–11170, 2010. doi: 10.1073/pnas.1005062107. Publisher: National Acad Sciences.

[4] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356, 2016. doi: 10.1038/nn.4244. Publisher: Nature Publishing Group.

[5] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv preprint*, 2018. doi: 10.1101/407007.

[6] Radoslaw Martin Cichy, Gemma Roig, Alex Andonian, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Yalda Mohsenzadeh, Kandan Ramakrishnan, and Aude Oliva. The algonauts project: A platform for communication between the sciences of biological and artificial intelligence. *arXiv preprint arXiv:1905.05675*, 2019.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. arXiv preprint arXiv:2002.05709.

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. doi: 10.1109/ICCV48922.2021.00951.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and others. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021. tex.organization: PMLR arXiv preprint arXiv:2103.00020.

[11] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Jascha Achterberg, Joshua B Tenenbaum, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

[12] Vighnesh Subramaniam, Colin Conwell, Christopher Wang, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Revealing vision-language integration in the brain with multimodal networks. *ArXiv*, pages arXiv–2406, 2024.

[13] Yena Han, Tomaso A Poggio, and Brian Cheung. System identification of neural systems: If we got it right, would we know? In *International conference on machine learning*, pages 12430–12444. PMLR, 2023.

[14] Jacob S Prince, Colin Conwell, George A Alvarez, and Talia Konkle. A case for sparse positive alignment of neural systems. In *ICLR 2024 Workshop on Representational Alignment*, 2024.

[15] David Colaço, Russell A Poldrack, Aliya Rumana, Thierault Jordan, Daniel C Burnston, Sofie Valk, Philipp Haueis, Daniel Marguiles, and Carl F Craver. Do dnns explain the visual system? guidelines for a better debate about explanation. 2025.

[16] Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lindsay, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, pages 1–20, 2023.

[17] Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, 15(1):9383, 2024.

[18] SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3):031003, 2018.

[19] SueYeon Chung and LF Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current opinion in neurobiology*, 70:137–144, 2021.

[20] Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119 (43):e2200800119, 2022.

[21] Bruno A Olshausen, David J Field, and others. Sparse coding of natural images produces localized, oriented, bandpass receptive fields. *Submitted to Nature. Available electronically as ftp://redwood. psych. cornell. edu/pub/papers/sparse-coding. ps*, 1995. Publisher: Citeseer.

[22] Cory Stephenson, Jenelle Feather, Suchismita Padhy, Oguz Elibol, Hanlin Tang, Josh McDermott, and SueYeon Chung. Untangling in invariant speech recognition. In *Advances in Neural Information Processing Systems*, pages 14368–14378, 2019.

[23] Abdulkadir Canatar, Jenelle Feather, Albert Wakhloo, and SueYeon Chung. A spectral theory of neural prediction and alignment. *Advances in Neural Information Processing Systems*, 36: 47052–47080, 2023.

[24] Eric Elmoznino and Michael F Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLoS computational biology*, 20(1):e1011792, 2024.

[25] Michael Kuoch, Chi-Ning Chou, Nikhil Parthasarathy, Joel Dapello, James J. DiCarlo, Haim Sompolinsky, and SueYeon Chung. Probing biological and artificial neural networks with task-dependent neural manifolds. In Yuejie Chi, Gintare Karolina Dziugaite, Qing Qu, Atlas Wang Wang, and Zhihui Zhu, editors, *Conference on Parsimony and Learning*, volume 234 of *Proceedings of Machine Learning Research*, pages 395–418. PMLR, 03–06 Jan 2024. URL `https://proceedings.mlr.press/v234/kuoch24a.html`.

[26] Ghislain St-Yves, Kendrick Kay, and Thomas Naselaris. Variation in the geometry of concept manifolds across human visual cortex. *bioRxiv*, pages 2024–11, 2024.

[27] Zhenan Shao, Yiqing Zhou, and Diane M Beck. Human visual robustness emerges from manifold disentanglement in the ventral visual stream. *Journal of Vision*, 25(9):1967–1967, 2025.

[28] Nancy Kanwisher, Meenakshi Khosla, and Katharina Dobs. Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, 2023.

[29] Emily Jean Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Logan T Dowdle, Bradley Caron, Franco Pestilli, Ian Charest, J Benjamin Hutchinson, Thomas Naselaris, et al. A massive 7t fmri dataset to bridge cognitive and computational neuroscience. *bioRxiv*, 2021.

[30] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. doi: 10.1109/CVPR.2009.5206848.

[31] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

[32] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[33] Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *Plos Biology*, 21(12):e3002366, 2023.

[34] Jenelle Feather, Meenakshi Khosla, N Murty, and Aran Nayebi. Brain-model evaluations need the neuroai turing test. *arXiv preprint arXiv:2502.16238*, 2025.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[36] Drew Linsley, Ivan F Rodriguez, Thomas Fel, Michael Arcaro, Saloni Sharma, Margaret Livingstone, and Thomas Serre. Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. *arXiv preprint arXiv:2306.03779*, 2023.

[37] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111. Publisher: National Acad Sciences.

[38] P. Bashivan, K. Kar, and J.J. DiCarlo. Neural population control via deep image synthesis. *Science*, 364, 2019. doi: 10.1126/science.aav9436. URL `https://doi.org/10.1126/science.aav9436`.

[39] Thomas Serre. Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science*, 5:399–426, 2019. doi: 10.1146/annurev-vision-091718-014951. Publisher: Annual Reviews.

[40] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, and others. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, 2019. doi: 10.1038/s41593-019-0520-2. Publisher: Nature Publishing Group US New York.

[41] Nathan Rose and Carlos R Ponce. Decoding face recognition abilities in the human brain. *Nature Communications*, 15:6371, 2024.

[42] Akshay V Jagadeesh and Justin L Gardner. Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences*, 119(17):e2115302119, 2022.

[43] Tyler Bonnen, Daniel LK Yamins, and Anthony D Wagner. When the ventral visual stream is not enough: A deep learning account of medial temporal lobe involvement in perception. *Neuron*, 109(17):2755–2766, 2021.

[44] Tyler Bonnen, Anthony D Wagner, and Daniel LK Yamins. Medial temporal cortex supports object perception by integrating over visuospatial sequences. *Cognition*, 262:106135, 2025.

[45] Fenil R Doshi and Talia Konkle. Cortical topographic motifs emerge in a self-organized map of object space. *Science Advances*, 9(25):eade8187, 2023.

[46] Johannes Mehrer, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, 2021.

[47] Jacob S Prince, Binxu Wang, Akshay V Jagadeesh, Thomas Fel, Emily Lo, George A Alvarez, Margaret S Livingstone, and Talia Konkle. Feature accentuation along the encoding axes of it neurons uncovers hidden differences in model-brain alignment. *Journal of Vision*, 25(9): 2215–2215, 2025.

[48] Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. VISSL, 2021. URL `https://github.com/facebookresearch/vissl`.

[49] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. URL `https://doi.org/10.5281/zenodo.5143773`.

[50] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[51] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. *Advances in Neural Information Processing Systems*, 35:8799–8810, 2022.

[52] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Toronto, Canada, July 2023. Association for Computational Linguistics. URL `https://aclanthology.org/2023.acl-demo.3`.

[53] Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, 13(1):1–12, 2022. doi: 10.1038/s41467-022-28091-4. Publisher: Nature Publishing Group.

[54] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018.

Appendix + Supplementary Information

Model Selection

Following the method of Conwell et al. [17], we curated a set of 117 deep neural network (DNN) models spanning different visual input diets (training data), architectures, and tasks. These models were sourced from the following repositories:

- the Torchvision model zoo [35];
- the VISSL model zoo [48];
- the DINO collection [9];
- the OpenAI CLIP collection [10];
- the OpenCLIP model zoo [49];
- the VicReg(-L) collections [50; 51];
- the Salesforce-LAVIS model zoo [52];
- the Open-IPCL collection [53];
- the Taskonomy model zoo [54];
- the SLIP model collection [**?**];
- and YOLOv5 [**?**]

The complete list of 117 models (organized by source) is shown below

Torchvision (53)
AlexNet-ImageNet1k-v1, ConvNext-Base-ImageNet1k-v1, ConvNext-Large-ImageNet1k-v1, ConvNext-Small-ImageNet1k-v1, ConvNext-Tiny-ImageNet1k-v1, DenseNet121-ImageNet1k-v1, EfficientNet-B1-ImageNet1k-v2, GoogleNet-ImageNet1k-v1, Inception-V3-ImageNet1k-v1, MobileNet-V2-ImageNet1k-v2, MobileNet-V3-Large-ImageNet1k-v1, MobileNet-V3-Large-ImageNet1k-v2, RegNet-X-1-6gf-ImageNet1k-v1, RegNet-X-3-2gf-ImageNet1k-v1, RegNet-X-3-2gf-ImageNet1k-v2, RegNet-X-400mf-ImageNet1k-v1, RegNet-X-400mf-ImageNet1k-v2, RegNet-X-800mf-ImageNet1k-v1, RegNet-X-800mf-ImageNet1k-v2, RegNet-X-8gf-ImageNet1k-v1, RegNet-X-8gf-ImageNet1k-v2, RegNet-Y-1-6gf-ImageNet1k-v1, RegNet-Y-16gf-ImageNet1k-v2, RegNet-Y-3-2gf-ImageNet1k-v1, RegNet-Y-3-2gf-ImageNet1k-v2, RegNet-Y-400mf-ImageNet1k-v1, RegNet-Y-800mf-ImageNet1k-v2, RegNet-Y-8gf-ImageNet1k-v1, RegNet-Y-8gf-ImageNet1k-v2, ResNet101-ImageNet1k-v1, ResNet152-ImageNet1k-v1, ResNet152-ImageNet1k-v2, ResNet18-ImageNet1k-v1, ResNet50-ImageNet1k-v1, ResNet50-ImageNet1k-v2, ResNext101-64x4d-ImageNet1k-v1, ResNext50-32x4d-ImageNet1k-v1, ShuffleNet-V2-X1-0-ImageNet1k-v1, ShuffleNet-V2-X2-0-ImageNet1k-v1, SqueezeNet1-0-ImageNet1k-v1, SSDLite320-MobileNet-V3-Large-Coco-v1, Swin-B-ImageNet1k-v1, VGG11-ImageNet1k-v1, VGG13-ImageNet1k-v1, VGG19-ImageNet1k-v1, ViT-B-16-ImageNet1k-v1, ViT-B-32-ImageNet1k-v1, ViT-L-16-ImageNet1k-v1, ViT-L-32-ImageNet1k-v1, Wide-ResNet101-2-ImageNet1k-v1, Wide-ResNet101-2-ImageNet1k-v2, Wide-ResNet50-2-ImageNet1k-v1, Wide-ResNet50-2-ImageNet1k-v2

VISSL (7)
ResNet50-Clusterfit, ResNet50-Deepclusterv2, ResNet50-Jigsaw-P100, ResNet50-MOCOv2, ResNet50-PIRL, ResNet50-SimCLR, ResNet50-SWAV

DINO (1)
ViT-B-16

CLIP (5)
RN50, RN101, ViT-B-16, ViT-B-32, ViT-L-14

OpenCLIP (8)
RN50-CC12m, RN50-YFCC15m, RN101-YFCC15m, Roberta-ViT-B-32-LAION2b-S12b-B32k, ViT-B-16-LAION400m-E32, ViT-B-32-LAION2b-S34b-B79k, ViT-B-32-LAION400m-E32, ViT-L-14-LAION2b-S32b-B82k

VicReg (2)
ResNet50, ResNet50-Alpha0p9

LAVIS (2)
ALBEF-Feature-Extractor-Base, Blip-Feature-Extractor-Base

VisionLab (5)
AlexNet-GN-IPCL-ImageNet1k, AlexNet-GN-IPCL-Mixedx3, AlexNet-GN-IPCL-OpenImagesv6, AlexNet-GN-IPCL-Places2, AlexNet-GN-IPCL-VGGface2

Taskonomy (22)
Class-Object, Class-Scene, Curvature, Denoising, Depth-Euclidean, Depth-Z-Buffer, Edge-Occlusion, Edge-Texture, Egomotion, Fixated-Pose, Inpainting, Jigsaw, Keypoints2d, Keypoints3d, Nonfixated-Pose, Normal, Point-Matching, Random-Weights, Room-Layout, Segment-Semantic, Segment-Unsup25d, Segment-Unsup2d, Vanishing-Point

SLIP (10)
ViT-B-CC12m, ViT-B-CLIP-CC12m, ViT-B-CLIP-YFCC15m, ViT-B-SimCLR-YFCC15m, ViT-B-

YFCC15m, ViT-S-CLIP-YFCC15m, ViT-S-SimCLR-YFCC15m, ViT-S-YFCC15m

YOLOv5 (3)
YOLOv5l, YOLOv5m, YOLOv5s