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

Anonymous Author(s)
Affiliation
Address
email

Abstract

What do deep neural network (DNN) models 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 – each 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 key 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 model comparison probes (e.g. category-supervised versus self-supervised models). 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 + Methods

High-level ventral visual cortex has often been considered the primary substrate of object recognition in the human brain [1; 2]. This notion has been further reinforced in recent years by the seminal finding that deep neural network (DNN) models supporting image classification are the most predictive models of ventral visual brain activity to date [3]. One lingering issue with this formulation, however, is the under-specification of what we mean by the word "recognition" in the context of the biological brain. What *precise* computational mechanisms instantiate this process? And how might we use deep neural network models to elucidate them? In deep neural network models, the computational mechanisms of 'recognition' are explicitly (i.e. algorithmically) defined. Canonical, 'end-to-end' image classification models (e.g. AlexNet [4]) '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 [5]) and masked decoding models (e.g. DINO) 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

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

classifiers (e.g. CLIP [6]) 'recognize' images by way of a top-k similarity score between the language embeddings of category labels and the vision embeddings of candidate images.

Identifying which of these (and many other) motifs is most brain-like has been somewhat confounded by the fact that many (if not most) of these models often achieve roughly comparable measures of 'representational alignment' [7] to human visual cortex as measured by neuroimaging [8] and neurophysiology [9] alike. In this work, we attempt to better parse the often difficult-to-interpret differences between better and worse models of the brain by comparing them to the scalar metrics of representational structure derived from manifold geometry analysis [10]. In so doing, we attempt as well to understand how this structure relates to the function of object recognition, and what our neural network may or may be telling us about the nature of this structure in the visual brain.

Model & metric curation: Our general approach was to first curate N=12 metrics of manifold geometry curated from two distinct lines of work: N=4 metrics from Chung et al. [11] (derived from replica mean field theory) and N=8 metrics from Sorscher et al. [12] (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. [13]. **Model-brain-predictivity**: Model-to-brain alignment (henceforth, brain-predictivity) 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 [14] 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. [11] and Sorscher et al. [12] were computed using N=50 'concept manifolds' (the representational matrices of 50 images from each of 50 category labels from ImageNet-1K/21K [15; 16], or Places365 [17]). To ensure proper evaluation of these metrics, we deployed the exact GitHub source code associated with the defining works of each: schung039/neural-manifolds-replicaMFT [11; 18] and bsorsch/geometry-fewshot-learning) [12].

2 Results + Analysis

Primary Analysis: Explaining Brain-Predictivity with Manifold Geoemetry: 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 [18]; results from this analysis are shown in Appendix Figure 1C and Table 3.

6 / 12 metrics showed significant rank-order correlation with brain-predictivity at $p < 0.01$. 3/6 of these metrics (manifold signal-to-noise ratio, capacity, and signal) showed significant, positive correlations. 3/6 (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. [12], 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. [11], at ρ = -0.724 [-0.857, -0.623], and the positively correlating manifold capacity metric from Chung et al. [11], at ρ = 0.616 [0.493, 0.779]. Notable as well in these correlations is a conceptual replication of previous results [13] that showed effective dimensionality (1 / 8 metrics from Sorscher et al. [19]) was not a significant predictor of brain-alignment – with ρ in this case equal to 0.156, p = 0.093.

The replication of this null effect, however, underscores all the more the significance of the positive effects evident in the *sizable* correlations of other metrics. And indeed, the primary takeaway from this first analysis might simply be that manifold metrics can indeed be the positive predictors of brain-alignment that theorists and empiricist alike have long proposed they might be [12; 20; 21; 22; 23].

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. [11]; Stephenson et al. [18]; Chung and Abbott [10] – 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. [12], 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.

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 [13], but seen also in our sample (see Figure 1C (left)). As also quantified with a segmented regression analysis pointing to rank 84 (ψ = 84.1 [72.4, 94.9]) as a notable breakpoint in an otherwise shallow slope of higher brain-predictivity scores, this set effectively included 84 models with average encoding scores of r_{Pearson} = 0.336 or higher. Here, already, manifold capacity begins to diffentiate from manifold SNR in terms of its rank-order correlation with brain-predictivity, with capacity's ρ diminishing to the point of non-significance at 0.14, p=0.222 and manifold SNR remaining subtantial and significant at 0.50, $p < .001$. In an even smaller subset of purely architectural variation (the N=53 category-supervised ImageNet-1K-trained models from the Torchvision model zoo [24]), the trend is similar, with ρ for manifold capacity dimishing 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. (More detailed results from this experiment are shown in Appendix Table 3.)

Experiments 2+3: 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 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 seems to be 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. [25]) to be strongly negative across our 53 models (ρ = -0.63, $p < .001$).

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. [18]), 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 $\rho = 0.457$ $p < .001$ 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.

(More detailed results from Experiments 1, 2, 3 are shown in Appendix Tables 3 and 3.

3 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 [26], the dominant, and in some ways most 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 [27; 28], this answer may be sufficient. For those seeking 'understanding', the search remains for other form of explanatory variables that supplement raw prediction with the parsimony of theories articulable in formal or natural language [29; 30; 23].

In this work, we attempt to instrumentalize tools from the emergent field of manifold geometry [10] in service of better understanding 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. First validating the second-order predictive power of manifold geometry metrics (i.e. the raw values of their rank-order correlation with brain-alignment), we found 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 found 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 seperability that still subserve recognition, but equally subserve the few-shot learning of new categories.

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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`.

[25] Drew Linsley, Ivan F Rodriguez 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. *Advances in Neural Information Processing Systems*, 36:28873–28891, 2023.

[26] 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.

[27] 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`.

[28] 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.

[29] 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.

[30] 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.

[31] Adrien Doerig, Rowan Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace Lindsay, Konrad Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, and others. The neuroconnectionist research programme. *arXiv preprint arXiv:2209.03718*, 2022.

[32] 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`.

[33] 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`.

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

[35] 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.

[36] 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`.

[37] 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.

Appendix + Supplementary Information

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.

Model Selection

Following the method of Conwell et al. [13], 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 [24];
- the VISSL model zoo [32];

	Manifold Capacity	Signal-to-Noise Ratio
	$\rho_{\text{Spearman}},\ p$	$\rho_{\text{Spearman}},\ p$
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.

	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.

- the DINO collection [33];
- the OpenAI CLIP collection [6];
- the OpenCLIP model zoo [33];
- the VicReg(-L) collections; [34; 35];
- the Salesforce-LAVIS model zoo [36];
- and Open-IPCL collection [37]

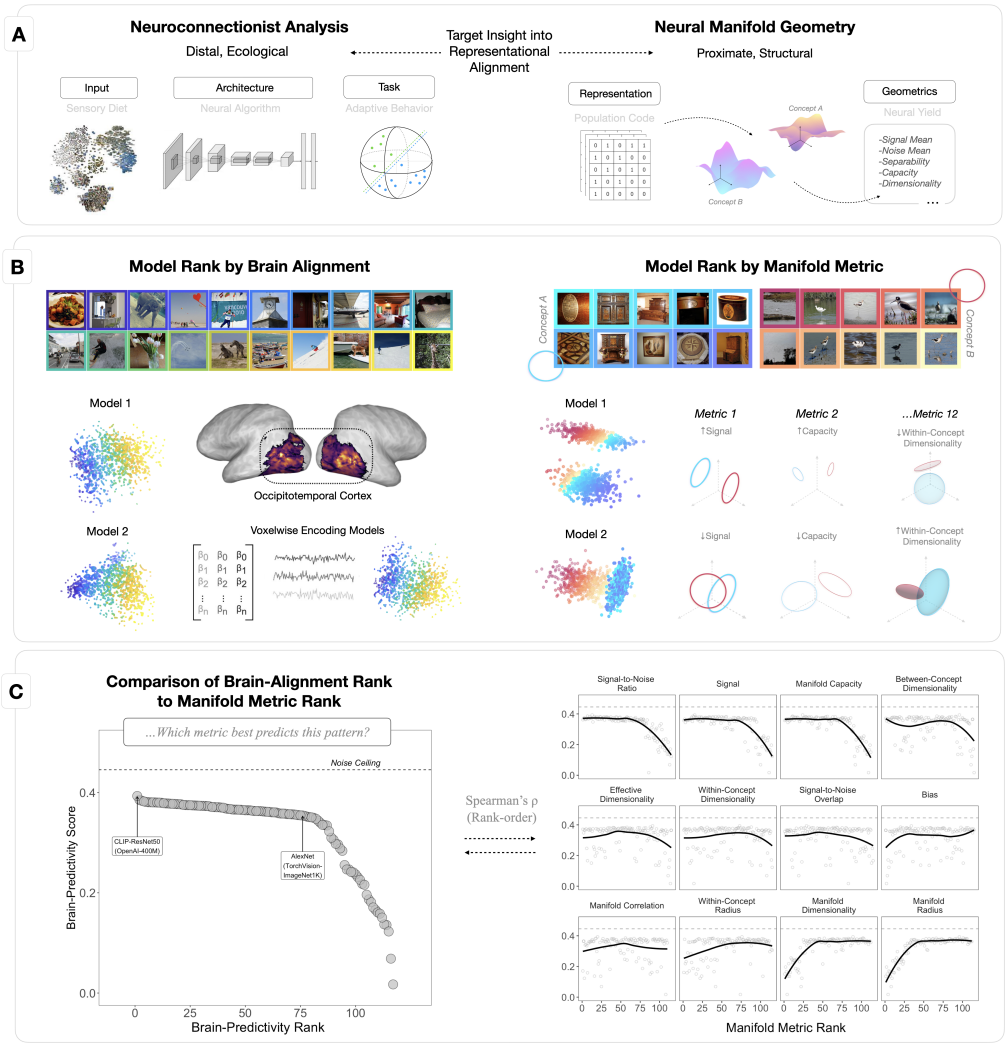

Figure 1: An overview of our motivation, methodology, and primary analysis. In **A**, we schematize the key factors that define and contrast the two animating frameworks of our model-to-brain comparisons. With its focus on the influence of input, architecture, and task, neuroconnectionist analysis [31] 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 [10; 12] 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-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' [14]. 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 the brain-predictivity on the x axis), which are sorted from top to bottom by Spearman's ρ.