Conclusions about Neural Network to Brain Alignment are Profoundly Impacted by the Similarity Measure

Ansh Soni[1*], Sudhanshu Srivastava[2], Konrad Kording[1,3], Meenakshi Khosla[2]

[1*]Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA.
[2]Department of Cognitive Science, UC San Diego, San Diego, CA, USA.
[3]Departments of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA.

*Corresponding author(s). E-mail(s): anshsoni@sas.upenn.edu;

Abstract

Deep neural networks are popular models of brain activity, and many studies ask which neural networks provide the best fit. To make such comparisons, the papers use similarity measures such as Linear Predictivity or Representational Similarity Analysis (RSA). It is often assumed that these measures yield comparable results, making their choice inconsequential, but is it? Here we ask if and how the choice of measure affects conclusions. We find that the choice of measure influences layer-area correspondence as well as the ranking of models. We explore how these choices impact prior conclusions about which neural networks are most "brain-like". Our results suggest that widely held conclusions regarding the relative alignment of different neural network models with brain activity have fragile foundations.

Keywords: NeuroAI

1 Introduction

Researchers have been building computational models to decipher neural activity for decades. While much of the earlier work focused on using hand-crafted stimulus features or behavioral variables to model brain activity, recent advances have seen deep neural networks (DNNs) emerge as powerful tools for understanding perceptual processing [1, 2]. These models have provided insights into the entire visual system, encompassing aspects such as eye movements [3], category selectivity [4, 5], and behavior on visual tasks [6, 7].

Early studies demonstrated that convolutional neural networks (CNNs) trained on behaviorally relevant tasks, like object categorization, exhibit representations remarkably similar to those found in the primate ventral visual stream [2, 8]. This alignment has led to the normative perspective that DNNs and the primate visual system may share common computational goals [9]. Inspired by these findings, researchers have increasingly focused on examining large collections of networks trained under various conditions to understand how different factors — such as architecture, training tasks, or learning rules — affect the alignment of their emergent representations (hidden layer activations) with those in biological systems [10]. The growing availability of large-scale neural activity data from diverse recording modalities (e.g., single-unit recordings, MEG, EEG, fMRI) has further enabled these extensive comparisons between models and brain data.

By modifying baseline models, researchers aim to determine what makes a model more brain-like, leading to the establishment of several large-scale benchmarks for identifying the most brain-like models. This approach has yielded significant insights into the effects of self-supervision versus category-supervision [11], the inclusion of language supervision [12, 13], adversarial training [14–16], and the enforcement of local correlations [17] on model-brain alignment.

Given the widespread interest in comparative analysis, numerous representational comparison measures have been proposed to evaluate different models against brain activity data. However, individual studies typically employ a single measure (e.g., linear neural predictivity), and the implementation of these measures often varies. Different measures are justified by distinct rationales: for instance, linear predictivity and regression methods are valued for their direct prediction of brain activity, their ability to generate new data, and for their use of simple linear linking functions between network activations and brain responses that can be seen as approximating downstream area readouts. Representational Similarity Analysis (RSA) and similar measures, which take in pairs of activation/response vectors and output a representational similarity defined by the inner product of these vectors, are favored for their symmetry, their ability to capture the geometric structure of information, and their lack of free parameters, thus mitigating overfitting concerns. Occasionally, measures are justified ad hoc, with statements like “Metric one doesn’t show any difference between models, whereas our other measure does.”

Different measures have distinct theoretical underpinnings and potentially varying empirical associations. Consequently, assessments of the alignment between computational models and biological systems can critically depend on the chosen measures. Whether the choice of measure significantly impacts results needs verification. To date, a systematic empirical comparison of these representational comparison measures has been lacking.

In this paper, we investigate the effects of measure choice on the brain-similarity for individual and across many Deep Neural Network models, examining how these differences in analysis techniques influence conclusions in the field. We find that the choice of measure is indeed not immaterial, with some measures showing poor associations with each other in how they rank different models or different layers within a model for a brain region, and that depending on the choice, some conclusions of prominent studies would have been meaningfully different. Overall, these results suggest that the outcome of these systems comparisons can depend critically on the comparative approach employed, and suggest a need to develop more comprehensive comparative methodologies that can produce robust results and help adjudicate not just which model is most 'brain-like' (the degree of similarity) but also the nature of that similarity.

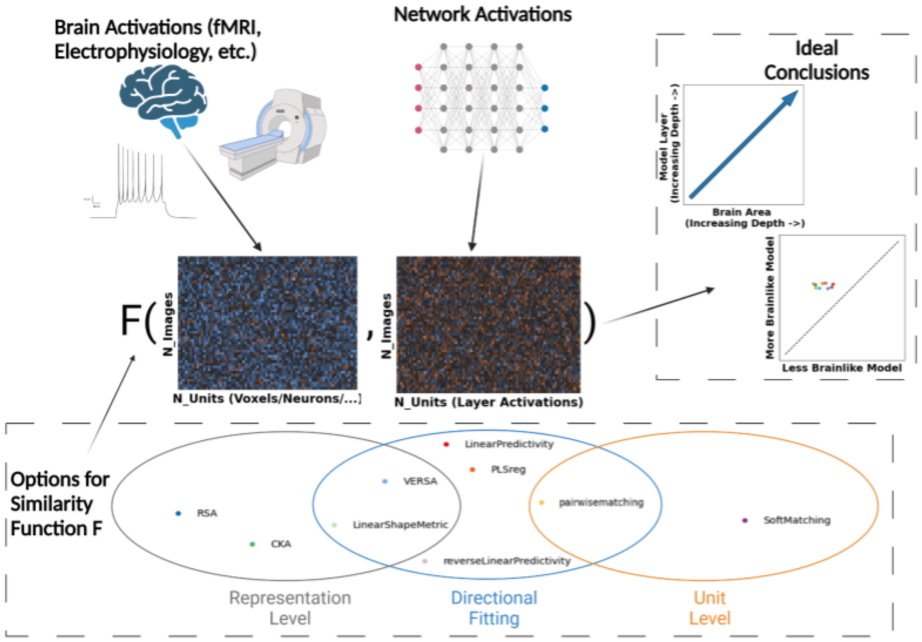
Brain Activations (fMRI, Electrophysiology, etc.)
Network Activations
Ideal Conclusions
Model Layer (Increasing Depth ->)
Brain Area (Increasing Depth ->)
More Brainlike Model
Less Brainlike Model
F(
N_Images
,
N_Images
)
N_Units (Voxels/Neurons/...]
N_Units (Layer Activations)
Options for Similarity Function F
LinearPredictivity
PLSreg
VERSA
RSA
LinearShapeMetric
CKA
pairwisematching
SoftMatching
reverseLinearPredictivity
Representation Level
Directional Fitting
Unit Level

Fig. 1 Asking how the choice of measure matters when comparing brain data with neural network activities. Activations from Brain and Model are extracted for a shared N_Stimuli utilizing various methods (e.g., single-unit recordings, MEG, EEG, fMRI). These activations are compared using a function that outputs a similarity score. Some choices for this function are listed (9 chosen for this paper). Measures have various theoretical similarities, and the main differences are highlighted. These similarity scores are then used to make various conclusions, such as which network is better and hierarchical correspondence. An ideal example is shown. Created with BioRender.com

Methods

We implement 9 different similarity measures. Each measure takes 2 matrices with a shared first dimension (N_Stimuli × N_Features) and outputs a similarity score. Non-Fitting measures such as representational similarity analysis (RSA [18]) and centered kernel alignment (CKA [19]) output a single score. Fitting measures like Linear Predicivity and voxel-encoded RSA (VERSA) output multiple scores as a result of a K-Fold validation which are then averaged for a final score. Comparisons are repeated for every subject within the dataset which are then averaged again. This gives 9 scores, one for each measure, for every dataset-model combination.

We also implement Sparse Random Projection as a possible dimensionality reduction step following the process of [10]. Comparisons with no dimensionality reduction were also completed.

Model activations were typically extracted after every block instead of every layer for comparison to brain data. For a more thorough comparison, we extracted activations for comparison after every layer for AlexNet.

We computed a ”maximum” through inter-subject similarity for each measure. We also computed 2 types of ”minimums”: one using the raw pixel values of the image, and another utilizing the category information of the images (the latter of which was only computed for the Natural Scenes Dataset (NSD) [20]).

Measure Implementation Details

Linear Predictivity

For Linear Predictivity, we fit a Ridge regression with the model neuron responses as the predictor variables, to predict voxel responses. The RidgeCV function in Scikit-learn was used with the regularization parameter, α, chosen as the optimal parameter from a logarithmic space between 10^{-8} and 10^{8}. The similarity score is then calculated as the R^2 on predicted responses to N unseen images.

Reverse Linear Predictivty

Reverse Linear Predictivity was calculated using the same procedure as Linear Predictivty, with the voxel responses being used as the predictor variables to predict model unit responses. Albeit not a commonly used measure, it has seen some use in the past [21].

Partial Least Squares Regression

Partial Least Squares regression was fit using the PLSRegression function in Scikit-learn with 25 components, to predict voxel responses using model neuron responses.

RSA

Representational Dissimilarity Matrices of size N by N were first computed for model responses as well as voxel responses by calculating the pairwise dissimilarity (1 - correlation) of each model neuron/voxel's responses to a pair of images. The similarity score was then computed using the Kendall Tau correlation between the upper triangle of the two matrices.

CKA

We used linear Centered Kernel Alignment (CKA), proposed in [19]. Linear CKA calculates the similarity between two matrices $X \in \mathbb{R}^{M \times N_x}$ and $Y \in \mathbb{R}^{M \times N_y}$ of neuronal responses using: $CKA(X,Y) = \frac{||X^TY||_F^2}{||X^TX||_F||Y^TY||_F}$ where F denotes the Frobenius norm. The matrices X and Y have centered columns.

VERSA

Voxel Encoded RSA follows the same ridge regression procedure as Linear predicitivity, the main difference is in calculating the score, instead of using the R^2, we create an RDM on the test set using the predicted values and use RSA to get the final score.

Linear Shape Metric

To compute the Linear shape metric, both the response matrices $X \in \mathbb{R}^{M \times N_x}$ and $Y \in \mathbb{R}^{M \times N_y}$ are first projected to a common dimension using PCA. Then, we compute the transformation T in the symmetry group G of Y that minimizes the distance between X and YT.

That is $d_{LSM}(x,y) = \min_{T \in G} ||X - YT||_F$, where $||.||_F$ denotes the Frobenius norm.

Both X and Y are mean-centered and normalized. This minimization is done via the procedure proposed in [22].

Pairwise matching

Pairwise matching compares two populations $X \in \mathbb{R}^{M \times N_x}$ and $Y \in \mathbb{R}^{M \times N_y}$ by first calculating, for each neuron in Y, the best matching neuron in X by calculating their response correlations on a training set of stimuli. Next, the correlation of each neuron in Y with its best-matching neuron in X is calculated on a testing set.

The procedure for calculating the pairwise matching is similar to the implementation presented in [7].

Soft Matching

As opposed to pairwise distance, Soft Matching is symmetric, meaning that the soft matching distance between X and Y is the same as that between Y and X.

Soft Matching is also implemented following the implementation presented in [7].

Models Used

AlexNet

AlexNet [23] is a small CNN, one of the first models in deep learning. We used the default PyTorch implementation of AlexNet pre-trained on ImageNet.

VGG

VGG [24] is a deeper CNN architecture with multiple depth options. We used the default PyTorch implementation of VGG networks pre-trained on ImageNet.

ResNet

ResNet [25] is a deep CNN with residual connections and multiple depth options. We used the default PyTorch implementations of ResNet networks pretrained on ImageNet.

Instance-Prototype Contrastive Learning (IPCL)

We used the pre-trained model weights from Konkle and Alvarez, 2022 [11].

Vision Transformers (ViT)

ViT is a transformer-based network [26]. We used the default PyTorch implementation of ViT pre-trained on ImageNet.

Topographic Deep Artificial Neural Network (TDANN)

We used the pre-trained model weights from Margalit et al., 2024 [17].

1.1 Variations in Metrics Over Many Papers

When choosing the implementations of each measure, we chose hyperparameters that best matched recent papers in the field. We keep this consistent throughout the paper, leading to small differences in measure hyperparameters between the comparisons presented here. For example, the 3 measures used in Margalit et al., 2024 [17] vary slightly from our implementation. First, their linear regression measure was a PLS regression with 1000 components we instead used 25 components. Second, when computing RSA, they collapsed across 5 categories, whereas we used the separate 515 images. Finally, our pairwise matching metric, while similar to their one-to-one metric, is slightly more permissive, allowing the same unit in one representation to be mapped to different units in another representation. If small changes in hyperparameters lead to large differences in comparisons, there should be a much larger focus on why these hyperparameters are selected.

Datasets Used

Natural Scenes Dataset

Detailed Descriptions of the NSD dataset and its collection can be found here [20]. Here is a brief summary. This dataset contains 7T fMRI data (1.8 mm, 1.6 s) for 8 participants, each viewing 9-10 thousand images multiple times totaling 22-30 thousand trials. Subjects completed a long-term continuous recognition task while viewing the images. We extracted ROI's from the data separating out V1,V2,V4, the rest of the Ventral stream (labelled VVS in plots), Dorsal stream (labelled DVS in plots), and Lateral stream (labelled LVS in plots). We used both the shared 515 images with 8 participants and shared 1000 images with 4 participants.

Object Orientation Dataset

Detailed Descriptions for this dataset can be found in Konkle & Alvarez 2021 [11]. Here is a brief summary. This dataset contains 3T fMRI data (1 mm,2.2s) for 4 participants, each viewing 40 images 4 times for a total of 160 trials. Subjects completed a vigilance task by pressing a button when a red circle appeared.

Inanimate Objects Dataset

Detailed Descriptions for this dataset can be found in Konkle & Alvarez 2021 [11]. Here is a brief summary. This dataset contains 4T fMRI data (3 mm,2.0s) for 8 participants, each viewing 72 images 6 times for a total of 432 trials. Subjects completed a vigilance task by pressing a button when a red frame appeared.

Monkey V1 Dataset

Detailed Descriptions for this dataset can be found in Cadena et al 2019 [27]. Here is a brief summary. Single neuron recordings to 7250 images for 2 monkeys. Stimuli were repeated 4 times. Due to more repeats than subjects and the low number of neurons (51 and 115), errors were taken across repeats as opposed to subjects.

ManyMonkeys Dataset

Detailed Descriptions for this dataset can be found in Dapello et al 2023 [16]. Here is a brief summary. These datasets contain single neural recordings to 640 images. Micro-electrode arrays were placed in the IT cortex of 6 monkeys leading to a varied number of 58-280 sited from each monkey. The firing rate was averaged over a 70-170 ms window following the onset of stimulus presentation.

Results

While there is Hierarchical Correspondence, it is Inconsistent Across Measures

One of the prominent early findings in the field of NeuroAI has been the ability of task-optimized DCNNs to replicate the hierarchical organization of the primate visual cortex [28, 29]. Specifically, different computational stages (layer depths) in these DNNs best align with different visual stages in the primate ventral visual cortex, mimicking their hierarchical progression. We find that this finding of early layers best aligning with early visual areas (V1, V2) and later layers best aligning with higher visual areas (IT) does seem to exist across many measures, although there is variance in which exact layer corresponds with which brain area (Fig. 1). In particular, the hierarchical correspondence breaks down for Soft Matching, pairwise matching, and Reverse Linear Predictivity, which seem to prefer either the same early convolutional layer or the same penultimate fully-connected layer. This finding is consistent with recent research on between-model comparisons, where soft matching was shown to be worse at clustering the same layer between multiple models with similar architecture and training [30] than linear predictivity and RSA. Far from being immaterial, the used measure affects the matching of layers to brain areas.

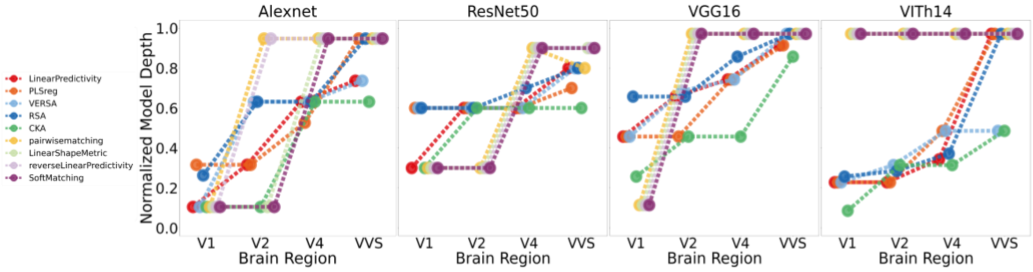

Fig. 2 Across measures, higher brain areas tend to best correspond to higher levels in the neural networks. Best fitting layer to brain regions (NSD Shared 1000, with Dimensionality Reduction) for four standard ImageNet trained models. Layer depth is normalized with 0 being the first layer and 1 the output layer. A slight jitter is added to the X-axis to make overlapping points more visible.

Measures Exhibit Weak Empirical Correlations

One way of obtaining an intuition about the relationship between measures is to look at their similarity across layers on a simple network. Using a popular example, we start by looking at between-metric correlations across layers on the network that started the DL revolution, AlexNet ([23]). If the choice of measure was immaterial, we would expect high correlations between the measures. However, the measures can be quite distinct, often having small or even negative correlations (Fig. 3 left). On Alexnet, the measures appear very distinct, suggesting that their choice may matter.

Not only the measure, but other methodological choices may also matter. For example, many comparison procedures start with a dimensionality reduction step to effectively deal with high-dimensional representations, which may also matter. Indeed, we find that the nature of dimensionality reduction has major impact on the similarity (Mean Kendall Tau 0.369, Fig. 2 right). The choice of dimensionality reduction has a major impact on the resulting measure similarity.

The dataset and the depth of the brain area recorded may also interact with the measure choice. We find again that both have a major impact on similarity (Mean Kendall Tau 0.395, Fig. 2 right). We see that the comparisons within the same dataset (Mean Kendall Tau 0.477) are more similar than the across datasets (Mean Kendall Tau 0.364) and find that the dataset has a greater impact than the depth of the brain area recorded ($p < 10^{-5}$). We see that multiple factors affect the relationships between measures including but not limited to dimensionality reduction, the dataset, and the depth of the recordings.

The metric similarity matrices do have some structure, and we see 2 major clusters of metrics with higher correlation arise. On one side, we have what seem to be more lenient metrics like linear predictivity and RSA, and on the other side, with stricter ones such as Pairwise Matching and Soft Matching. Some metrics also don't seem to belong in their group, for example, reverse linear predictivity is unexpectedly in the strict matching side as typically forward and reverse linear predictivity are thought to be similar. We comment more on this difference later in the section "Forward Vs. Reverse Linear Predictivity". This does allude to the possibility of utilizing different metrics based on the context, although a greater theoretical undertaking is required to specify the use cases.

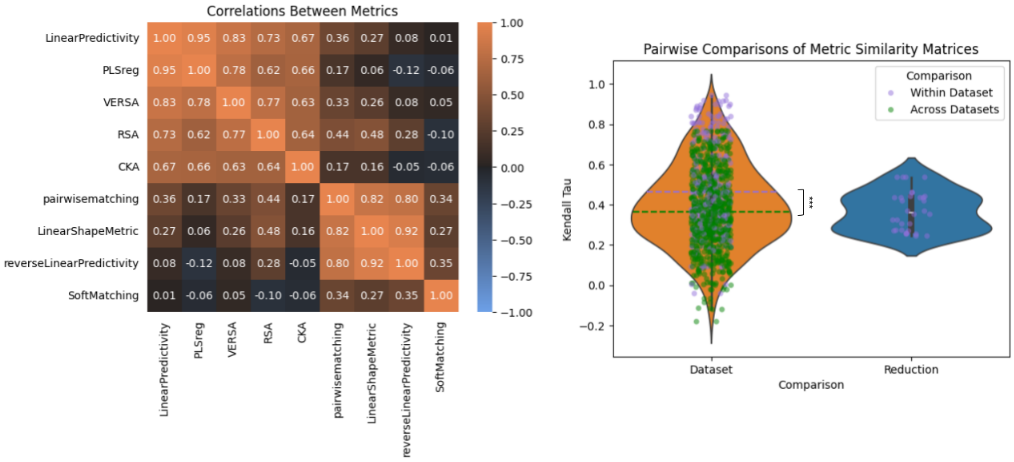

Fig. 3 There are significant differences between measures when compared on AlexNet Left: Mean Measure Similarity Matrix (MSM) of across layer correlation between measures. The mean is taken across datasets separated by brain area (16 options) and use of dimensionality reduction (2 options) for a total of 32 MSM's. The matrix is ordered by the correlation to Linear Predictivity. Right: Kendall Tau Values for pairwise comparisons of RSMs. In orange the dimensionality reduction is constant with comparisons between different datasets separated by brain area. In blue, the dataset is constant and comparison with or without dimensionality reduction. The comparisons where both RSM's are from the same dataset (but different brain areas) are plotted in green (ex. Both sides are from NSD with no dimensionality reduction but one is V1 and another is V2). The similarity of comparisons within dataset has a significantly higher similarity than the rest.

Different Measures Yield Varying Ranking of Models

The most common use of this comparative analysis approach is to draw conclusions about the brain by comparing how well different models – each representing different hypotheses about neural computations through distinct design and training choices – align with neural data. For the conclusions drawn to be robust, the alignment ranking between models must also be consistent across different comparative measures. We find that this is not the case. After finding the max alignment across layers for each individual model, comparing this ranking across models is also highly dependent on measure choice. Across 25 models, we see a large discrepancy in ranking so that a model that is least aligned for one measure can even be in the top half for another measure

(Fig. 4 top). We further see that the ranks on average exhibit low correlation across measures for V1 (mean correlation of 0.13 Fig. 4 bottom) with higher correlations in VVS (mean correlation of 0.43 Fig. 4 bottom), with certain groupings of measures have much higher correlation (similar to findings in Fig. 3). This low consistency in rankings will lead to different conclusions about the brain even when everything else is done in a controlled manner.

Forward Vs. Reverse Linear Predictivity

Typically forward and reverse linear predictivity are thought to be similar in ideal conditions. We instead find the 2 have low correlation for model rankings and lead to different conclusions. A few benign reasons can lead to this. For example, there is often a large difference in dimensionality between the model and brain data. Furthermore, the methods used to sample the brain data can be sparse (electrophysiology) or have significant degradation (fMRI). In our comparisons there are two ways to address these concerns. Firstly for the rank correlations (Fig. 4), dimensionality reduction is used so that the fMRI data has similar dimensionality (on the order of 10^3). The degradation in fMRI can also be addressed by using dimensionality reduction on the model side. This point is not fully mitigated as the degradation of information is not the same for the model and the brain. Still seeing these differences between the forward and reverse linear predictivity, even with attempts to control for mitigating circumstances, can lead to rethinking the types of conclusions made about brain model alignment utilizing directional fitting metrics.

How Much Do We Know?

Typically, the inter-subject similarity scores are used as a level of maximum predictability, with untrained models used as a baseline. Seeing that untrained models are not always worse than trained models we used the pixel values of the image, and the category information as two separate baselines. We see that depending on the measures and brain area there is a large variability between the measure and how well the average model performs according to the measure. Some measures show that the average model is aligned better to brain activity than other models while other measures show that the pixel or category values are better than the model itself. We want to emphasize here, that the models are compared to individual brains, making it well possible for fitting measures to predict activity better than the inter-subject test. In this sense, people often think of the model fit as comparable to the brain of another person, however, that way of thinking breaks down due to individual brain fitting.

Revisiting Conclusions

Following our finding of the inconsistency of measures, we revisit some prior conclusions in the space to see how they change depending on the chosen measures and brain areas.

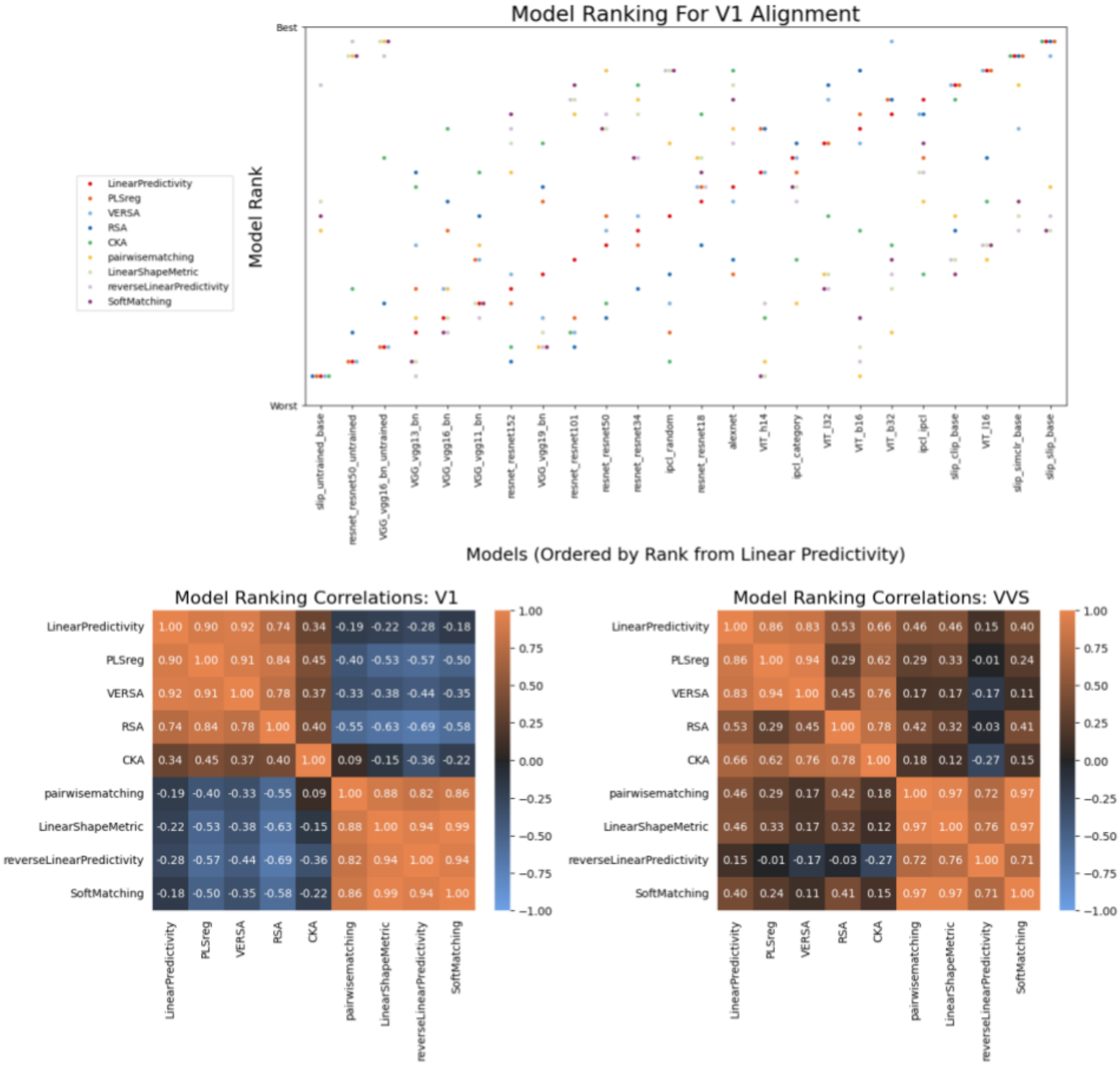

Fig. 4 The choice of Measure affects the preferred models. Top: Models ranked from worst to best alignment to V1 data (NSD Shared 1000, with Dimensionality Reduction) with Linear Predictivity. Rankings for other measures are shown with different colors. Dimensionality reduction is applied. Bottom: Representational Similarity Matrix (RSM) of Spearman correlation between measures for V1 and VVS.

Unsupervised vs. Supervised Learning

Scientists in NeuroAI ask if Unsupervised Learning vs Supervised learning produces networks that have better alignment to brain activations [11]. After all, if the brain learns in an unsupervised way, we may expect that unsupervised learning will produce

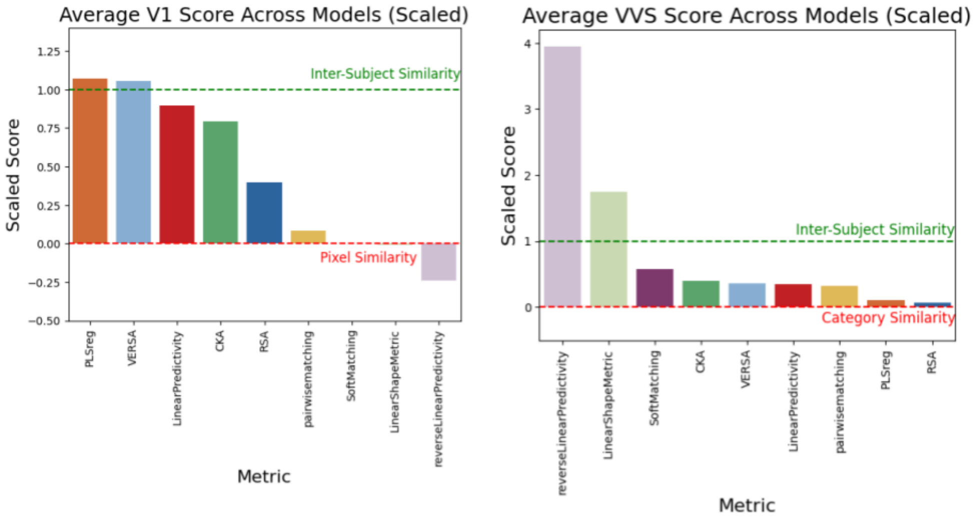

Fig. 5 Some measures seem misleadingly good while others appear disappointingly bad. The average similarity score (NSD Shared 1000, with Dimensionality Reduction) across many models for V1 (Left) and the post-V4 Ventral Visual Stream (Right). The value is scaled by the inter-subject alignment at 1 and either the pixel values for the image (Left) or the category information (Right) at 0 .

better models. The mechanisms proposed to do this are also biologically plausible as most unsupervised learning methods work through learning invariant embeddings for multiple versions of a single image and can, in turn, be compared to learning invariant percepts for multiple eye movements and perspectives of a given scene.

Utilizing the same dataset and models of [11] (due to the controlled training of their open-source models), we find conflicting results for category-supervision versus unsupervised learning. We see that there is a slight bias towards the unsupervised learning method, which aligns with the conclusions of the paper (6 Left). We see a more concerning trend, especially for EarlyV, where since we scale the values by subtracting the score of the untrained model, the values of measures below 0 indicate that they are performing worse than the untrained model. Almost a third of the measure/brain region comparisons indicate that the untrained model is doing better than both trained models and a half indicated it is doing better than at least one trained model. Our findings suggest that most of the brain alignment effects come from the architecture as opposed to the training but that the choice of measure matters more than the supervised vs unsupervised condition.

The Importance of Language

Scientists ask if networks that are multi-modally trained with visual and text information, arguably a more human-like dataset, have better alignment to Brain activations albeit with varying conclusions about the benefit of language [10, 12, 13, 31, 32].

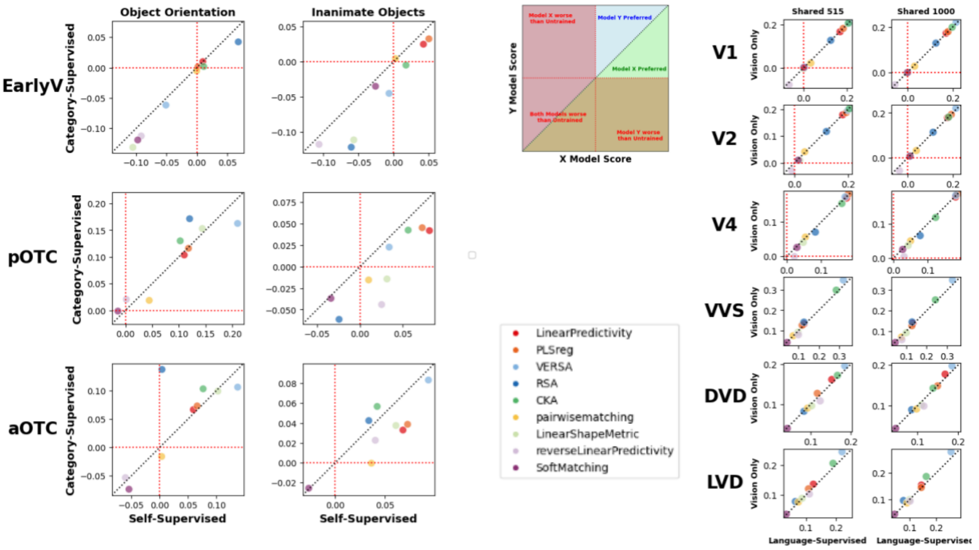

Fig. 6 Conclusions about Self-supervised models or multi-modal models fitting brain data better than baseline models are highly fragile to measure choice. Left: Unsupervised (IPCL) vs. Category Supervised Learning. Each plot shows the scores for each measure for 2 datasets (Object Orientation and Inanimate Objects, without dimensionality reduction) over 3 broader brain regions of increasing depth. Each score is scaled by the score of the untrained models' alignment. Scores under the dashed red lines mean the trained model has a lower score than the untrained model. Scores below the black diagonal prefer unsupervised models, and above prefer category supervised. Right: Each plot shows the scores for each measure over 6 brain regions of increasing depth (NSD, with dimensionality reduction). For the language-supervised side, the score of the better-aligned model between the Language-Only and Language+Vision models is chosen. Each score is scaled by the score of the untrained model's alignment. Scores under the dashed red lines mean the trained model has a lower score than the untrained model. Scores below the black diagonal mean that some level of language supervision improves alignment, and above means that language supervision hurts alignment.

Like a lot of those papers, we use SLIP models, a group of 3 models trained on a similar dataset with similar hyperparameters. The only differences are that one is trained with language supervision, one is trained with image-based self-supervision, and the last one is a combination of both. We see that while almost all of the measures still show the trained models are better than the untrained model, there is little to no difference between training methods with most measures on diagonal and equal off-diagonal towards each side. Deeper brain areas do seem to be more off-diagonal, although there isn't a strong bias toward either side. We see that strong conclusions about the importance of multi-modal training appear problematic, the results strongly depend on the choice of the measure: while some measures lead us to prefer multi-modal models, others lead us to prefer single modality models.

Topographic Networks

Scientists also ask if having similar topography in the models as the brain improves the alignment to brain activations [17]. They introduced an innovative technique to regulate the amount of spatial correlations in the neural networks they train. This excitingly replicates spatial patterns of cortex, such as the pinwheel organization in V1, similar to prior work based on independent component analysis [33]. On top of the exciting set of findings about spatial organization, such as pinwheels existing in the model, they made two surprising discoveries that make use of a measure. (1) They find that when they regulate the spatial coherence leading to topography similar to that of the human brain (through topographic metrics), it at the same time leads to the best model-brain alignment in VTC. (2) They also find that their chosen self-supervised learning objective combined with that level of spatial correlation reaches close to ceiling levels of alignment in VTC. They concluded that setting the right correlation topography produces models with known spatial structure but also models that make sense from an alignment perspective.

To ask if the choice of measures matters we reanalyzed their modeling setup with some minor methodological differences to keep the analyses of this paper coherent (PLS regression with 25 instead of 1000 components, RSA across images instead of categories, and a different implementation of pairwise matching. See Methods). We find that the measure, again, has a strong influence on the resulting model brain alignment scores (Fig. 7).

We see that only one metric satisfies both (1) the alpha dependence and (2) the training objective dependence. This pairwise matching metric is also closest to the metric that shows both in the paper. A closer reading of the paper shows us that this choice is justified by the inability of their PLS regression metric to separate the models. While we commend the explanation's existence, it also highlights a potential problem in the underlying methodology. With a range of metrics and even more hyperparameters within each metric, similar justifications could be used to exemplify differences in models that are not really there. The possibility of a measure not only influencing the alignment score itself but also the scientific claims downstream of the measure should bring a much higher level of scrutiny on the method.

Discussion

Comparing representations in biological and artificial neural networks has become a widely used analytical approach for understanding the computational logic and mechanistic functioning of various brain regions. In this rapidly expanding ecosystem of model-brain comparisons, numerous tools and measures have emerged to facilitate these comparisons. However, most papers creating guidelines or criticizing are not empirically grounded [34–38], and there is still a lack of empirical understanding regarding how these measures relate to one another in terms of yielding similar conclusions. This gap means that the choice of measures or tools in different studies is often driven by tradition, lacking a principled basis and sufficient justification.

Several studies have tested current model evaluation tools by benchmarking popular similarity measures based on various theoretical and empirical criteria across

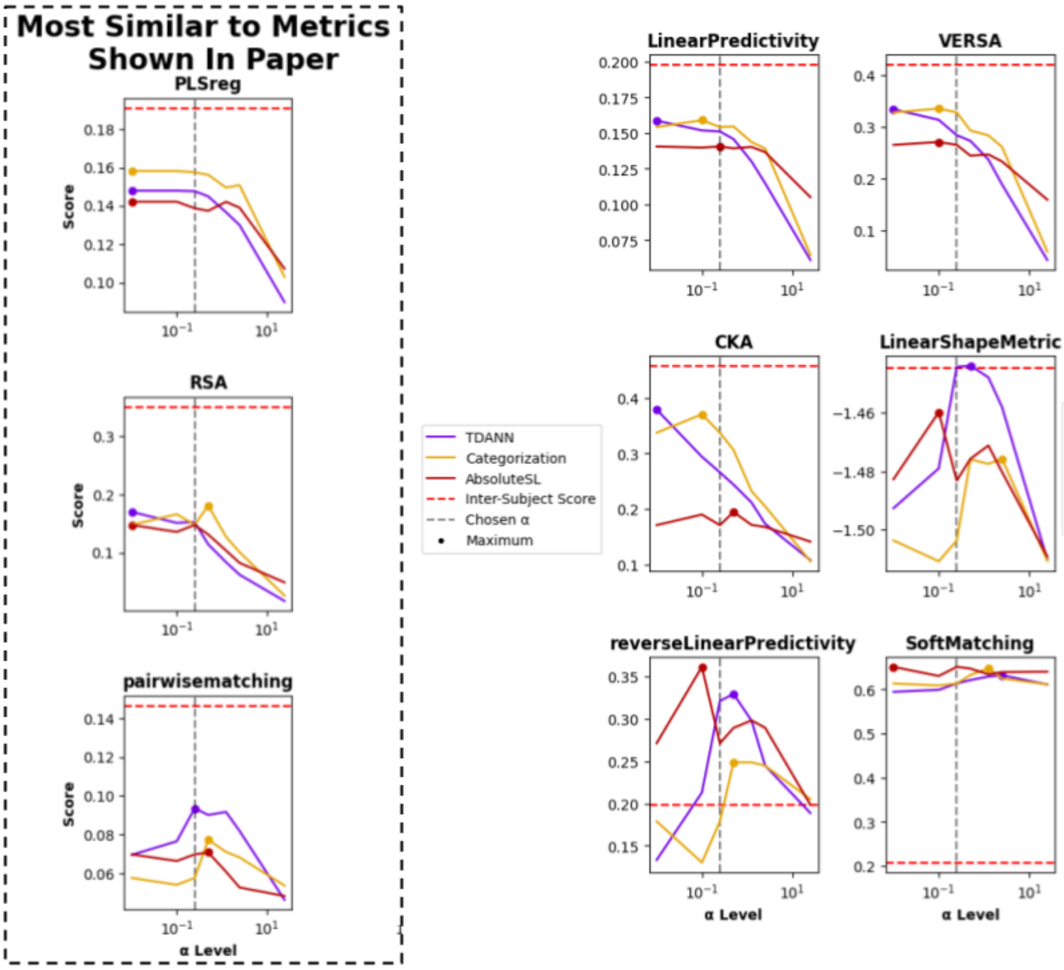

Fig. 7 Matching the spatial structure of brains only makes models more brainlike for a limited choices of measures. Replotting Figure 6 (B) from Margalit 2024 [17] with 9 measures (NSD VVS without Dimensionality Reduction). The dots represent which level of spatial correlation leads to the best alignment for each learning rule. The gray line denotes the Paper's chosen value that matches topographic properties of the brain such as pinwheels

different neural networks. The theoretical criteria consider the invariants that measures should obey, while the empirical criteria focus on the desired outcomes, specifically, some assumed similarity structure of neural network representations, that ideal measures should produce. For instance, Han et al. [39] assessed different similarity measures based on their effectiveness in system identification, i.e., their ability to identify the ground truth (a model with the same architecture as the target system) among various candidate models. Kornblith et al. [19] explored the use of different similarity measures in yielding reliable correspondences between the layers of different artificial neural network models and proposed that an ideal similarity measure should be invariant to orthogonal transformations and isotropic scaling but not to invertible linear transformations and found that centered kernel alignment (CKA) best met these empirical and theoretical criteria. Ding et al. [40] examined the sensitivity of methods like canonical correlation analysis (CCA), CKA, and orthogonal Procrustes distance to changes in factors that do not affect the functional behavior of neural network models. However, these studies compare only a few measures, primarily focusing on classical measures like RSA, CKA, and encoding analysis. While there are clearly theoretical reasons to prefer some measures over another, we should consider the breadth of used measures until the field has agreed on one.

Here we comprehensively evaluated widely used measures across multiple datasets and models to examine the relationships between these measures and the extent to which different measures support similar conclusions. We have shown that not only does the measure cause inconsistencies in the brain alignment within and across models but these inconsistencies undermine some of the conclusions made in the neuroAI field. The effects of unsupervised learning and language supervision in producing more brain-like models remain unclear, with different measures yielding different conclusions, sometimes even showing low effect sizes when compared to untrained networks. If distinct measures, that are all defendable, produce wildly different conclusions about the relationship between brain representations and different learning rules, architectures and data sources, then these findings cannot be considered robust. This discrepancy raises the question of how we can develop analysis methods that will make our results more reliable. On a more optimistic note, we do see some more consistent conclusions, such as the hierarchical correspondence, which, albeit noisy, seems to be a consistent trend across measures. However, some measures, specifically the ones that seem to emphasize a more strict mapping of units, seem to consistently favor either the first highest dimensional layer or the final lowest dimensional layer for all brain areas, without much room in between.

It is also important to keep in mind that we are unable to test all of the hyperparameters in completing these comparisons. For example the metric used to compare the RDM's in RSA can significantly vary. We used Kendall Tau while papers in the literature have used rank correlation or even CKA. The LinearShapeMetric has a tuning parameter alpha that can vary between 0 and 1 (we use alpha=1 as recommended in the initial paper). There are also hyperparameters in which model activations to use. Some papers use the best fitting layers, others use only the final embedding layer, and some others even use multiple layers to compare to a single brain region. Such a wide range of hyperparameters for multiple aspects of brain-model comparisons leads to significant concern about manipulation. While this isn't required to be the case, it does mean we should make these decisions carefully and look upon them with more scrutiny.

Due to the complex multi-scale relations within brains, our comparison of measures does not contain a deeper statistical analysis. The difficulty here is that statistical methods for analyzing such data are complicated. We could obtain confidence intervals for the measures through bootstrapping, but along with the bulk of the field, we omitted them here. This omission makes sense when relatively large datasets are involved. It also makes sense, given how the confidence intervals are affected by many aspects of the data. Future analyses of model-brain alignment should have more meaningful ways of evaluating modeling results.

Our analysis treated all measures as being indistinguishable. However, there may be places with a-priori reasons to prefer one measure over all others. Some of the previously mentioned work for inter-model similarity could be considered in this case, among other papers investigating measures in relation to the brain [30, 41]. In those cases, our analysis that the measure matters is still a true observation, but it should not detract from the results of the paper. After all, there was, arguably, only one good measure to choose. However, a careful reading of the papers in this field suggests that the measures were often not chosen specifically from an analysis of the problem at hand but chosen for compatibility with the wider field or mathematical elegance. As such, there are real concerns about the choice of measure, in particular, given the incentives for coming up with conclusions compatible with the beliefs and intuitions of the field.

We also note that the inconsistency between measures in yielding different conclusions about various models is not necessarily a flaw but can be intriguing if interpreted correctly. Claims should emphasize not only the degree of similarity between two representations but also the nature of that similarity. Different measures capture different notions of correspondence based on their invariances and sensitivity to various neural representational properties, such as geometry, information content, and representational form and hence produce different results.

Using different measures together may provide complementary insights into the nature of similarity between representations. Additionally, such comparisons with multiple measures offer an opportunity to revisit and critically analyze the implicit assumptions of existing representation comparison techniques, leading to the development of novel tools to address their limitations. For example, the observed discrepancies between linear predictivity (neural response variance explained by models) and reverse linear predictivity (model response variance explained by neurons) in this paper highlight the limitations of unidirectional measures. However, considerable work may be needed to produce interpretable insights from distinct result obtained by using multiple measures.

We note that, in addition to the measures studied in this paper, there are other tools available for assessing representational similarity. For example, some suggest using interventional measures that go beyond correlations to evaluate how substituting one representation with another affects downstream computations of the network [42], using measures that do not assume a Euclidean representational space [43], as well as measures that quantify the hierarchical brain-like structure of a network [44]. We leave a thorough comparison of these tools with other measures analyzed in this study for future work.

A recent paper acknowledged the large variability in findings across a few measures on their specific setting [45]. Under Linear predictivity, their untrained model exhibited the maximum brain alignment, while CKA and RSA led to different conclusions. They also noted some unexpected effects of dataset shuffling and trained vs. untrained networks that are aligned with the findings of this paper. The choice of measure clearly is an emerging problem.

We want to briefly dwell on our thankfulness to the field. All the specific claims in this paper are chosen due to their recent importance, relevance to the field, and most importantly the availability of public models. These models have in common that they are controlled apart from single manipulations, having one tuned hyperparameter. We stress the importance of two points: it is extremely important to publish full data and code and any claims relying solely on similarity scores needs to be revisited with a closer examination of the models. The openness of the NeuroAI field allows meta-analyses like ours.

NeuroAI is an emerging field. As such, it is not overly surprising that we have to figure out the tools we use to study the alignment between brains and models. Relative to most past research in neuroscience, alignment problems are very high-dimensional and the choice of a fitting measure is influenced by the many distinct strategies available in high-dimensional spaces. Our findings on the significance of measures underscore the unique challenges in this field and emphasize the need to adapt our methodologies to handle the complexity of the data we are modeling.

Code Availibility

A GitHub repository with the data and code to reproduce all figures and experiments is provided here: github.com/anshksoni/NeuroAIMetrics

References

[1] Schrimpf, M. *et al.* Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* (2020). URL https://www.cell.com/neuron/fulltext/S0896-6273(20)30605-X.

[2] Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Computational Biology* (2014). URL https://doi.org/10.1371/journal.pcbi.1003915.

[3] Kümmerer, M., Bethge, M. & Wallis, T. S. Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision* **22**, 7–7 (2022).

[4] Khosla, M. & Wehbe, L. High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv* 2022–03 (2022).

[5] Ratan Murty, N. A., Bashivan, P., Abate, A., DiCarlo, J. J. & Kanwisher, N. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature communications* **12**, 5540 (2021).

[6] Prince, J. S., Alvarez, G. A. & Konkle, T. A contrastive coding account of category selectivity in the ventral visual stream. *bioRxiv* 2023–08 (2023).

[7] Khosla, M., Williams, A. H., McDermott, J. & Kanwisher, N. Privileged representational axes in biological and artificial neural networks. *bioRxiv* 2024–06 (2024).

[8] Yamins, D. L. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences* **111**, 8619–8624 (2014).

[9] Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience* **19**, 356–365 (2016).

[10] Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A. & Konkle, T. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv* 2022–03 (2022).

[11] Konkle, T. & Alvarez, G. A. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications* **13**, 491 (2022).

[12] Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J. & Wehbe, L. Incorporating natural language into vision models improves prediction and understanding of higher visual cortex. *BioRxiv* 2022–09 (2022).

[13] Subramaniam, V. *et al.* Revealing vision-language integration in the brain with multimodal networks. *arXiv preprint arXiv:2406.14481* (2024).

[14] Kong, N. C., Margalit, E., Gardner, J. L. & Norcia, A. M. Increasing neural network robustness improves match to macaque v1 eigenspectrum, spatial frequency preference and predictivity. *PLOS Computational Biology* **18**, e1009739 (2022).

[15] Dapello, J. *et al.* Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *Advances in Neural Information Processing Systems* **33**, 13073–13087 (2020).

[16] Dapello, J. *et al.* Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness (2023). URL https://openreview.net/forum?id=SMYdcXjJh1q.

[17] Margalit, E. *et al.* A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron* (2024).

[18] Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience* **2**, 249 (2008).

[19] Kornblith, S., Norouzi, M., Lee, H. & Hinton, G. *Similarity of neural network representations revisited*, 3519–3529 (PMLR, 2019).

[20] Allen, E. J. *et al.* A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience* **25**, 116–126 (2022).

[21] Higgins, I. *et al.* Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications* **12**, 6456

(2021).

[22] Williams, A. H., Kunz, E., Kornblith, S. & Linderman, S. W. *Generalized shape metrics on neural representations*, Vol. 34 (2021).

[23] Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks **25** (2012). URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[24] Simonyan, K. & Zisserman, A. *Very deep convolutional networks for large-scale image recognition* (2015).

[25] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. arxiv e-prints. *arXiv preprint arXiv:1512.03385* **10** (2015).

[26] Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[27] Cadena, S. A. *et al.* Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology* **15**, e1006897 (2019).

[28] Güçlü, U. & van Gerven, M. A. J. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience* **35**, 10005–10014 (2015). URL https://www.jneurosci.org/content/35/27/10005.

[29] Cichy, R., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. (2016). URL https://doi.org/10.1038/srep27755.

[30] Thobani, I., Sagastuy-Brena, J., Nayebi, A., Cao, R. & Yamins, D. L. *Inter-animal transforms as a guide to model-brain comparison.*

[31] Doerig, A. *et al.* Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737* (2022).

[32] Conwell, C., Prince, J. S., Hamblin, C. J. & Alvarez, G. A. *Controlled assessment of clip-style language-aligned vision models in prediction of brain & behavioral data* (2023).

[33] Hyvärinen, A., Hoyer, P. O. & Inki, M. Topographic independent component analysis. *Neural computation* **13**, 1527–1558 (2001).

[34] Cao, R. & Yamins, D. Explanatory models in neuroscience, part 1: Taking mechanistic abstraction seriously. *Cognitive Systems Research* 101244 (2024).

[35] Cao, R. & Yamins, D. Explanatory models in neuroscience, part 2: Functional intelligibility and the contravariance principle. *Cognitive Systems Research* **85**, 101200 (2024).

[36] Bowers, J. S. *et al.* Deep problems with neural network models of human vision. *Behavioral and Brain Sciences* **46**, e385 (2023).

[37] Kanwisher, N., Khosla, M. & Dobs, K. Using artificial neural networks to ask 'why'questions of minds and brains. *Trends in Neurosciences* **46**, 240–254 (2023).

[38] Doerig, A. *et al.* The neuroconnectionist research programme. *Nature Reviews Neuroscience* **24**, 431–450 (2023).

[39] Han, Y., Poggio, T. A. & Cheung, B. *System identification of neural systems: If we got it right, would we know?*, 12430–12444 (PMLR, 2023).

[40] Ding, F., Denain, J.-S. & Steinhardt, J. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems* **34**, 1556–1568 (2021).

[41] Acosta, F., Conwell, C., Sanborn, S., Klindt, D. A. & Miolane, N. *Evaluation of representational similarity scores across human visual cortex* (2023).

[42] Sexton, N. J. & Love, B. C. Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science advances* **8**, eabm2219 (2022).

[43] Shahbazi, M., Shirali, A., Aghajan, H. & Nili, H. Using distance on the riemannian manifold to compare representations in brain and in models. *NeuroImage* **239**, 118271 (2021).

[44] Nonaka, S., Majima, K., Aoki, S. C. & Kamitani, Y. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *IScience* **24** (2021).

[45] AlKhamissi, B., Tuckute, G., Bosselut, A. & Schrimpf, M. Brain-like language processing via a shallow untrained multihead attention network. *arXiv preprint arXiv:2406.15109* (2024).