ONLY BRAINS ALIGN WITH BRAINS: CROSS-REGION PATTERNS EXPOSE LIMITS OF NORMATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Normative models of brain regions aim to replicate their representational geometry and are widely used to study neural computation. Model–brain alignment is typically assessed with metrics such as representational similarity analysis (RSA) and linear predictivity (LP). Recent studies, however, show that conclusions from such benchmarks depend strongly on the choice of metric, raising a deeper conceptual problem: What do we truly mean with "brain alignment"? We address this by testing a broad spectrum of vision models on the BOLD-Moments video fMRI dataset and analyzing the influence of the alignment metric in greater detail. While benchmarks can identify a nominally best model, many other models fall within subject-level variability and are therefore practically equivalent. To move beyond metric dependence, we introduce Alignment Pattern Similarity (APS), a framework that uses brain-to-brain alignment as ground truth for evaluating normative models. For each region, we compare its empirical alignment with other regions against the alignment obtained when replacing that region's activity with its normative model. Strikingly, while normative models can align well with their target region, their cross-region alignment patterns diverge systematically from those observed in the brain. This reveals a key deficiency: current normative models do not faithfully reproduce brain-to-brain alignment patterns when substituted for real neural data. Furthermore, we show that structural connectivity can predict aspects of these alignment patterns, illustrating how anatomical constraints may additionally guide expectations about functional correspondence. Overall, APS shows great promise to become a principled framework for more robust and biologically meaningful assessments of brain–model alignment.

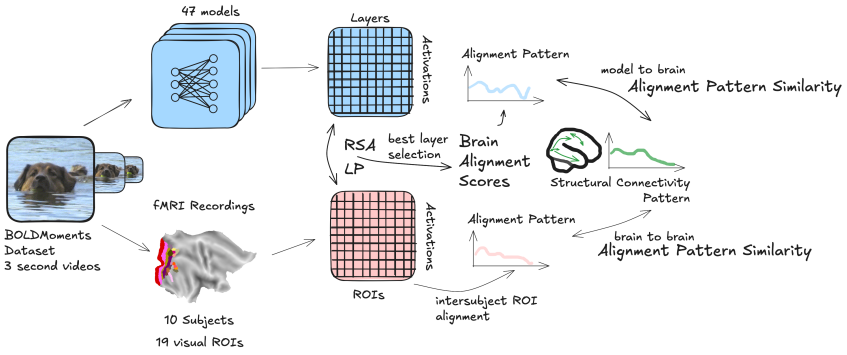

Figure 1: **Overview of our methodology**: Using the BOLDMoments dataset (Lahner et al., 2024), we benchmark model–brain alignment in visual cortex ROIs by comparing fMRI responses with model activations via linear predictivity (LP) and RSA. Many models appear practically equivalent under these measures, motivating a stricter metric—Alignment Pattern Similarity (APS)—which assesses how well similarity patterns (across brains and between models and brains) align with anatomical evidence of ROI–ROI relationships.

1 INTRODUCTION

A central aim of computational neuroscience is to uncover the functions performed by neural populations. For visual processing, decades of research have resulted in a putative mapping of distinct tasks such as object recognition and motion processing onto hierarchical pathways across visual cortex (Goodale & Milner, 1992; Kravitz et al., 2011; Rolls, 2024). Yet it remains unclear which evolutionary and developmental inductive biases shape the primate visual system in a way that supports robust visual behaviors.

Over the last ten years, researchers have leveraged increasingly powerful artificial vision models to probe candidate inductive biases. In this line of research, task-driven deep learning models are evaluated with regard to how well they *align* to functional recordings from visual brain areas (Yamins & DiCarlo, 2016). Alignment is assessed using a variety of metrics (Klabunde et al., 2025), and high alignment is typically interpreted as evidence that a model's task, architecture, and training dataset reflect biologically relevant constraints. Early work demonstrated that networks trained for object recognition predict responses in high-level ventral visual cortex (e.g. Yamins et al., 2014), laying the foundation for investigating core object recognition as the primary task of the ventral visual stream (e.g. Schrimpf et al., 2018). Recent progress in computer vision, including the advent of large-scale self-supervised and multimodal models (e.g. Bear et al., 2023; Assran et al., 2025a), along with the systematic collection of large-scale fMRI datasets (Lahner et al., 2024; Allen et al., 2022) have further broadened this research program. These developments have enabled comprehensive benchmarks of vision models in terms of their brain alignment.

The conclusions drawn from such benchmarks, however, rest on the assumption that alignment metrics measure brain–model alignment in a meaningful way. Recent work has challenged this assumption, showing that commonly used metrics can yield inconsistent results and may not capture alignment in a principled way (Bowers et al., 2022; Schaeffer et al., 2024; Soni et al., 2024). This raises a deeper question: *What does it actually mean for a model to be "brain-aligned"?*

Recent benchmarks of video models have relied on a single alignment metric, while metric-comparison studies rarely include modern video architectures — leaving it unclear how much conclusions for state-of-the-art vision models depend on metric choice. Here, we evaluate state-of-the-art vision models against large-scale fMRI data from the BOLDMoments dataset (Lahner et al., 2024), shifting the focus from single-metric rankings to the stability of rankings and their consistency with biologically meaningful brain connectivity.

In summary, our work makes the following contributions:

- We benchmark state-of-the-art vision models on the BOLDMoments dataset with both RSA and LP, showing that large-scale video models (e.g., V-JEPA2, Assran et al. (2025b)) achieve the strongest brain alignment across brain areas.
- We move beyond simple rankings by quantifying the robustness of model orderings, showing that many models are effectively equivalent within subject-level variability; RSA provides modest gains in discriminability over LP.
- We introduce *Alignment Pattern Similarity (APS)*, a connectivity-grounded criterion revealing that brain–brain similarity follows anatomical connectivity, while model–brain alignment remains weak.

Taken together, our findings call for stricter, biologically grounded criteria for assessing brain–model alignment, and we discuss below how they motivate changes to current benchmarking pipelines.

2 RELATED WORK

Alignment benchmarks. Kicked off by work on explaining visual object recognition (Yamins et al., 2014), neural network models have been compared to the brain on large-scale benchmarks. Brain-Score (Schrimpf et al., 2018) originally focused on static image processing along the ventral stream, later adding language regions (Schrimpf et al., 2021). It provided a first large-scale platform for model evaluation. The Algonauts challenge brought this approach to whole-brain responses to naturalistic stimuli, beginning with static images and later extending to dynamic movie-based

paradigms (Gifford et al., 2024; 2023; Cichy et al., 2021). Most recently, the challenge has emphasized multimodal video inputs, pushing alignment analyses into richer and more ecologically valid contexts (see e.g. d'Ascoli et al., 2025).

Beyond these community benchmarks, several studies have systematically examined factors shaping model–brain alignment. Conwell et al. (2024) showed that vastly different architectures can achieve similar alignment, with variation in "visual diet" emerging as the most consistent determinant. Tang et al. (2025) further found that a single predictive objective generalizes well across cortical areas when evaluated with LP, suggesting shared computational principles across the hierarchy. In contrast, Sartzetaki & Groen (2025) used RSA to reveal stream-specific alignment: modular architectures preferentially aligned with dorsal versus ventral pathways, consistent with a division of labor between motion and object processing. Sartzetaki et al. (2024) find that while video models achieve highest RSA-alignment in early visual regions, for both ventral and dorsal regions, semantic objectives seem key.

Alignment metrics. Conclusions about brain–model alignment strongly depend on the choice of alignment metric. Several recent studies have shown that different metrics can yield inconsistent model rankings, highlighting the instability of current benchmarking practices (Soni et al., 2024; Bo et al., 2024). In particular, LP has drawn substantial criticism: Schaeffer et al. (2024) argue that LP primarily reveals biases of the regression framework rather than genuine alignment, while Soni et al. (2024); Bo et al. (2024); Wu et al. (2025) show that LP offers low discriminability between models. More broadly, Bowers et al. (2022) contend that such metrics do not provide a reliable basis for drawing conclusions about brain alignment at all. Together, these findings underscore the need for stricter and more interpretable criteria when assessing model–brain correspondence.

3 METHODS

3.1 DATASET

We base our analyses on the BOLDMoments dataset (Lahner et al., 2024), a 3T fMRI dataset recorded from 10 subjects watching over 1000 different 3-second video clips. We chose this dataset to ensure stimulation of motion-responsive brain areas (Grossman & Blake, 2002; Sunaert et al., 1999). Each of the 998 stimuli in the train split was shown three times to each subject, each of the 98 stimuli in the test split was shown ten times. Stimulus repetitions were presented in random order across 4 sessions. We use beta values (GLMSingle regression coefficients of each voxel and video shown), projected to fslr32k surface space, as they are output from the preprocessing pipeline (specifically, version B) of Lahner et al. (2024). Please refer to Lahner et al. (2024) for more details. We use the original train-test split, and average the fMRI activity over repetitions, leading to a higher signal-to-noise ratio on the test split, compared to the train split. While the voxel-wise beta values provided by Lahner et al. (2024) are already centered and normalized across individual sessions, we normalize and center them once more across the train set, and use the same standard deviation and mean per feature to approximately center and normalize the test split.

We analyze ROIs from the Glasser HCP-MMP atlas (Glasser et al., 2016): early visual areas (V1,V2,V3), dorsal stream (V3A, V3B, V6, V6A, V7, IPS1, MST, MT, FST, LO1–LO3), and ventral stream (V4, V8, PIT, FFC).

3.1.1 NOISE-CEILINGS AND ROI-ROI SIMILARITY BASED ON INTER-SUBJECT RELIABILITY

Because fMRI data are noisy, neither perfect predictivity nor perfect representational similarity can be expected (Walther et al., 2016). We compute noise ceilings and similarities between different ROIs in the following way, for each subject: We average the fMRI data of the remaining N-1 subjects for the first, 'source' ROI. Then we use this average as the predictor feature space for the two metrics and compute the RSA/LP score between this average map and the left-out subject's ROI data for the second 'target' ROI. For noise ceilings, 'source' and 'target' are the same ROI. We do not normalize results by the noise ceiling but additionally display the ceiling where helpful.

3.2 MODELS

We evaluate 47 state-of-the-art pretrained image and video deep learning models that cover a broad range of architectures, objectives and datatsets:

Taskonomy model bank. A collection of 26 models based on ResNet-50 and trained on the same dataset of 4 million indoor scenes, but for different tasks (Sax et al., 2018; Zamir et al., 2018). **Supervised image models.** We include ResNet (He et al., 2016) and ConvNext (Liu et al., 2022b) models from the timm library (Wightman, 2019), all trained for object recognition on ImageNet-1K. **Self-supervised image models.** As counterpart to the supervised image models, we include ResNets trained on ImageNet-1K with the self-supervised SimCLR objective (Chen et al., 2020), as provided by VISSL (Goyal et al., 2021). **CLIP.** We consider the ResNet-50 and Vision Transformer based CLIP models from the original codebase, all trained to align image and text representation on a large dataset of 400M image-text pairs (Radford et al., 2021). **Supervised video models.** We use three video transformers from the mmaction2 toolkit (Contributors, 2020): MViT (Li et al., 2022), Video Swin Transformer (Liu et al., 2022a), TimeSformer (Bertasius et al., 2021). All models were trained for action recognition on the Kinetics-400 dataset. **Unsupervised video models.** We include the ViT-based counterfactual world model (CWM) (Stojanov et al., 2025) which was trained on Kinetics-400 using an adapted MAE objective (He et al., 2022). Further, we consider the V-JEPA 2 model (Assran et al., 2025a) that was trained on a large-scale video dataset using a variant of the MAE objective in feature space. **VGG Transformer (VGG-T).** We include the 3D foundation model VGG-T (Wang et al., 2025) as comparison to the dominantly semantic models described above. This ViT-based model was trained to simultaneously predict multiple key 3D attributes from a variable number of views of a scene.

For all models, we extract representations for the last layer of up to 15 blocks (e.g., a residual blocks in a ResNet). For models with more blocks, we use 15 equally spaced blocks. We apply image models to each frame individually, and video models and VGG-T to the entire video clip (3s, and average representations over time. The resulting feature vectors are reduced using sparse random projection (Achlioptas, 2003) to 5919 dimensions, following the Johnson-Lindenstrauss Lemma with an epsilon of 0.1 (Achlioptas, 2001).

3.3 MEASURING MODEL-BRAIN ALIGNMENT

For every combination of model and ROI, we select the best layer on the training set by averaging the alignment scores over subjects. Using the selected layer for all subjects, we then report alignment scores on the test set. We consider the following two alignment metrics:

Representational Similarity Analysis (RSA) compares representations based on representational dissimilarity matrices (RDMs), which are sufficient statistics for the representational geometry of a system (Kriegeskorte & Wei, 2021; Kriegeskorte et al., 2008). RDMs are constructed for the model and brain representation by computing the pairwise correlation distances of the representation ($1-$correlation) for all samples. The overall RSA alignment score is then the correlation of the brain and model RDMs.

Linear predictivity (LP) measures alignment by fitting a linear model that predicts brain activity from model features (e.g., Yamins et al. (2014)). We fit ridge regression models predicting the preprocessed fMRI signals on the training set using 5-fold cross-validation. We use the RidgeCV implementation from the scikit-learn package (Pedregosa et al., 2011), which selects the optimal alpha value using leave-one-out cross-validation from 19 candidate values on a logarithmic scale spanning 10^{-9} to 10^{9}. Given the respective linear models fitted on the training set, we report the residual sum of squares (R^2) on the test set.

3.4 DETERMINING PRACTICAL EQUIVALENCE BETWEEN MODELS

To assess when models can be considered practically indistinguishable in terms of brain alignment, we defined an equivalence criterion based on subject-level variability. Variability was estimated via bootstrap resampling: subjects ($n = 10$) were sampled with replacement, mean ROI-level alignment was recomputed for each resample, and 95% confidence intervals were derived from the resulting distribution. A model was deemed practically equivalent to the top-ranking model if its average alignment score fell within the 95% confidence interval of the top model.

3.5 CONNECTIVITY-BASED ALIGNMENT PATTERN ANALYSIS

For any given ROI, we define its alignment pattern as the vectorof similarities between this ROI and all ROIs (incl. the ROI in question) analysed. The similarity s can be given by a metric such as RSA and LP. We additionally define a metric- and data-independent measure of similarity based on anatomical connectivity. Specifically, we arrange ROIs in an undirected graph $G = (V, E)$, where each ROI corresponds to a vertex v and there exists an edge $e_{i,j}$ between two vertices v_i, v_j if there is a known feedforward connection between these two ROIs (Rolls, 2024). Note that we do not take feedback connections into account here. For the connectivity-based alignment similarity between ROI i and j, we then calculate the length l of the shortest path between v_i, v_j, and set the similarity to $s_c(i, j) = s_c(j, i) = \lambda^l$, with $\lambda = 0.9$. The alignment pattern of ROI i then results as the vector $S_c(i) = \{s_c(i, j)\}$ for $j = 1, ..., N_{rois}$. Metric-based alignment patterns derived from functional (i.e. fMRI-measurements) are defined analogously, with $s_{RSA}(i, j)$ simply given by the RSA score between ROI i and j. Alignment pattern similarity is then defined as the Pearson correlation between $S_c(i)$ and $S_{RSA/LP}(i)$. For model-to-brain comparisons, $s_{RSA/LP}(m, j)$ is the similarity/predictivity between model m and ROI j. APS is then computed between highly scoring models for ROI i, $m_1^i, ...m_K$ as the correlation between $S_{RSA/LP}(m_k^i)$ and $S_c(i)$.

4 RESULTS

In the following sections, we first present the overall results regarding the alignment of models and neural data (4.1). We then analyze the influence of the respective metrics on the results in greater detail (4.2 and 4.3), and conclude by exposing the limits of SOTA model brain alignment by comparing to brain-brain alignment (*Alignment Pattern Similary (APS)*, 4.4).

4.1 BENCHMARKING ALIGNMENT OF VISION MODELS TO THE VISUAL CORTEX

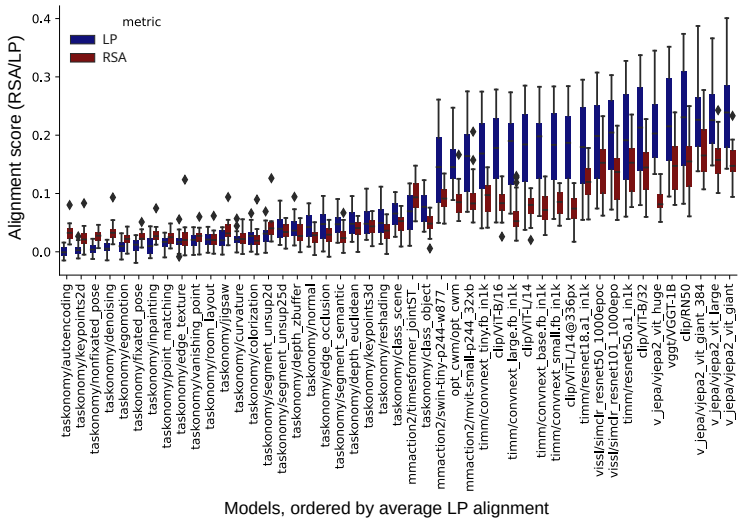

Figure 2: Alignment scores across models. Boxplots depict the distribution of subject-averaged alignment scores (RSA/LP) across ROIs, with models ordered by their average LP alignment.

We evaluated a range of vision models on their alignment to visual cortex—including early, ventral, and dorsal regions (Fig. 2)—using two complementary alignment metrics: RSA and LP. The models varied with respect to architecture (CNNs and Transformers), training objective (various supervised and self-supervised objectives), modality (image and video), as well as model size and training dataset (Methods 3.2).

Consistent with previous work (Tang et al., 2025), we found that the self-supervised V-JEPA 2 model family (Assran et al., 2025a), achieved the strongest overall alignment across visual cortex, according to both RSA and LP. Notably, however, the best aligned models included CLIP with a

ResNet-50 backbone and the VGG-Transformer—which differ in several important aspects from vJEPA and each other, such as the training data, training objective and overall architecture.

LP appeared to primarily separate poorly aligned models (largely from the Taskonomy family) from the rest, while offering limited discrimination among better-aligned models. RSA, by contrast, produced a more graded ranking that distinguished among high-performing models.

4.2 Practical equivalence in model-brain alignment

Next, we further assessed the robustness of model rankings and the discriminability between models in terms of brain alignment as predicted by both metrics. To this end, we examined alignment at the level of individual regions-of-interest (ROIs), checking for models that were practically equivalent in brain-alignment (see Methods 3.4). This analysis revealed that many models were practically equivalent in terms of brain alignment (Fig. 3, Fig. S11). For example, in early visual cortex (V1), LP grouped eight models as practically equivalent to the best model, whereas RSA reduced this set to just two. A similar pattern emerged in ventral regions such as PIT, where LP identified ten models as equivalent compared to five with RSA. Dorsal regions, by contrast, showed broader equivalence classes under both metrics. Overall, RSA yielded sharper distinctions than LP, classifying on average only 4.1 models as practically equivalent per region, compared to 5.73 with LP (Fig. S8).

These findings demonstrate that ranking models without considering the magnitude of their alignment differences can be misleading, since many models are often practically equivalent—underscoring the importance of incorporating stricter or additional measures when evaluating brain–model alignment.

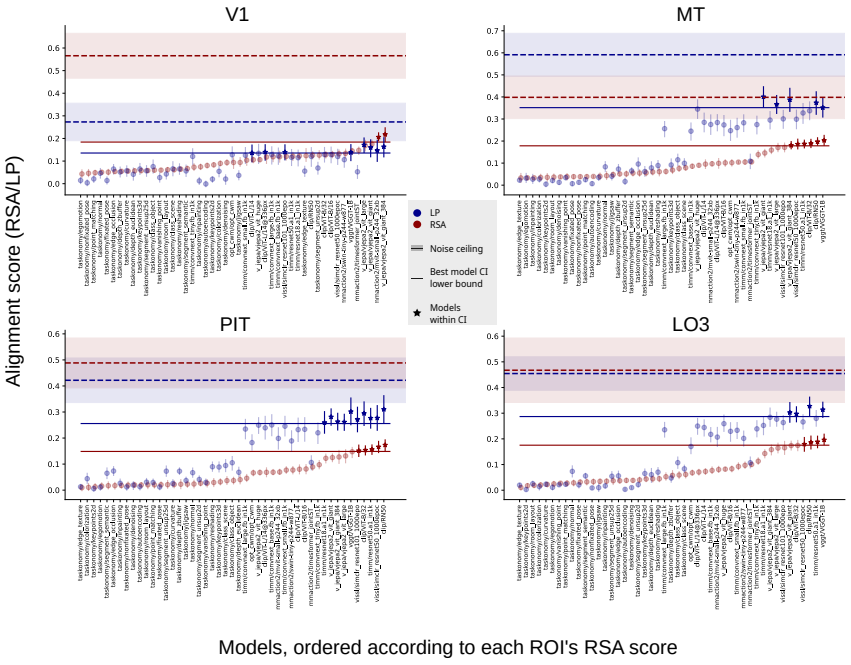

Figure 3: Model rankings for individual ROIs. Models are ordered according to each ROI's RSA alignment score. Opaque symbols denote models outside the 95% confidence interval (CI) of the most-aligned model, while star-highlighted symbols indicate models within the 95% CI. Dotted lines show the inter-subject noise ceiling, and horizontal bars mark the lower bound of the 95% CI for the top model.

4.3 Different modeling approaches yield equivalent brain alignment

Building on the results from the previous sections, we evaluated whether models that are practically equivalent share major design parameters, such as model architecture, training dataset or training objective.

Again, we find that this depends on the metric and also on the brain region being analyzed (Fig. 4). For instance, RSA identified unsupervised transformer architectures trained on video datasets as top performers in all early visual regions as well as V4, the entry point of the ventral stream. This pattern shifted in later ventral regions: In FFC, image-based supervised models—both CNNs and transformers—ranked highest. Dorsal regions showed a more heterogeneous picture, but a similar trend emerged, with unsupervised video models dominating earlier areas and supervised image models performing better in later ones.

Conversely, LP yielded a less differentiated picture. Across regions, it largely identified the same set of high-performing models, but tended not to exclude models that RSA ranked lower. More generally, rather than highlighting clear shifts between model types across the hierarchy, LP often grouped a broader range of architectures, objectives, and modalities as equivalently aligned.

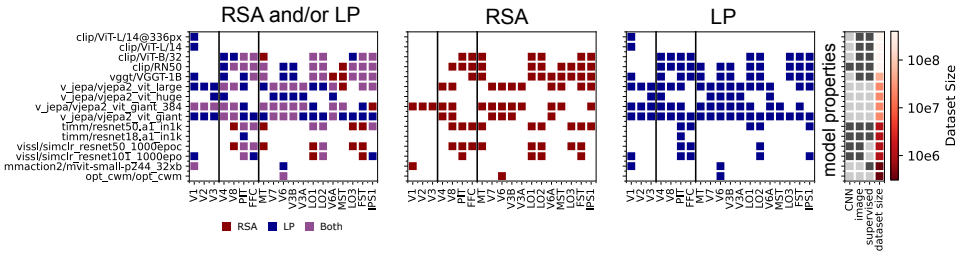

Figure 4: Equivalently brain-aligned models for RSA and LP. The 2D plots show models (y-axis) versus ROIs, with colored markers indicating equivalently aligned models per ROI: LP (blue), RSA (red), and overlap (purple). The right panel summarizes properties of the models along the y-axis, including dataset size used for training.

In summary, while both metrics distinguish poorly aligned models from better aligned ones, LP offers little discriminability among the large set of well-performing models. RSA separates these models more effectively and highlights region-specific preferences, yet even RSA sometimes assigns similar alignment to models that differ in architecture, modality, training objective, or dataset (e.g., in V8). This raises a broader question: What does it actually mean for a model to be "brain-aligned"?

4.4 Brain alignment as alignment pattern similarity

We propose that a model is *meaningfully aligned* to a brain region if it not only matches that region locally, but also preserves its pattern of relationships to other regions. For example, V1 is most similar to V2, less similar to V3, and very different from PIT. A V1-aligned model should exhibit the same relational profile. We refer to this relational profile as a region's *alignment pattern*.

We define alignment patterns independently of any functional alignment metric. Otherwise, a model might appear aligned simply because it shares an idiosyncratic feature of that metric, rather than capturing the functional profile of the region. Anatomical connectivity provides a suitable proxy because it constrains and shapes functional interactions across the visual system: Regions that are strongly connected are more likely to share information and thus exhibit similar representational profiles. Based on recent literature on the human connectome (Rolls, 2024), we constructed a connectivity graph (Fig. 5a, see Methods 3.5) and defined similarity between regions as a function of the length of the shortest path between them. We included ROIs along the well-established ventrolateral and dorsal stream with sufficiently clear connectivity. This yielded a connectivity-based similarity matrix (Fig. 5b), where each row or column corresponds to a region's alignment pattern (Fig. 5c).

We validate this approach by predicting brain-to-brain functional alignment patterns between ROIs as measured by RSA and LP via cross-subject prediction: For each ROI, the average activity of $N-1$ subjects served as the predictor feature space, and the activity of the held-out subject as the target (Fig. 5d). We correlated (Pearson) each ROI's connectivity-based alignment pattern with its functional alignment pattern, yielding an *alignment pattern similarity* (APS) score (Fig. 5d, right). Both metrics produced relatively high APS at the group level, reaching up to 0.88 (RSA) and 0.82 (LP) for V1, with mean values of 0.5 += 0.23 (RSA) and 0.65 += 0.15 (LP) across ROIs.

Different from high brain-to-brain alignment, we find that models do not achieve high APS. For each ROI, we compute the alignment pattern of the top-ranked (Fig. 5e, stars) and the practically equivalent models (Fig. 5e, boxplots; see Methods 3.4) under each metric and compared it to the ROI's connectivity-based alignment pattern. Strikingly, both metrics sometimes selected models with poor or even negative APS (Fig. 5e). For example, RSA's top model for V1 (V-JEPA2 ViT-Giant 384) achieved an APS of 0.63, whereas LP's top model (V-JEPA2 ViT-Large) scored –0.31. Generally, RSA tended to favor models with higher APS in early and ventral regions, while the pattern was more mixed in dorsal areas. Overall, the models selected by the benchmarking pipeline using RSA as alignment metric achieve an average APS of 0.24, whereas the models selected by the pipeline using LP as alignment metric achieve an average APS of 0.02.

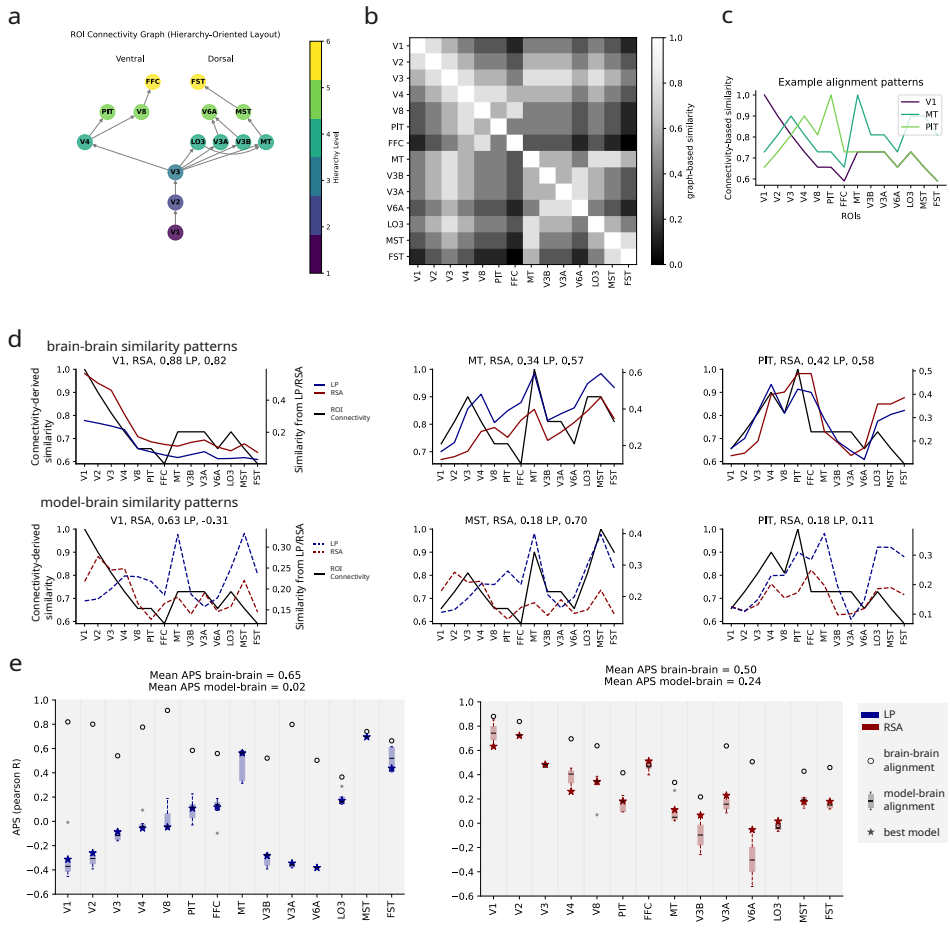

Figure 5: Alignment pattern similarity across ROIs and models. a) Connectivity graph defining the similarity structure based on known connectivity (Rolls, 2024). b) Corresponding similarity matrix derived from the graph. c) Example alignment patterns for three ROIs (V1, MT, PIT) relative to all others. d) Top: brain–brain similarity patterns for RSA and LP compared to the connectivity-derived pattern. Bottom: similarity patterns of the most ROI-aligned models for RSA and LP, for three example ROIs. e) APS for best-performing (star) and equivalent (boxplots) models for LP (left) and RSA (right).

These results indicate that current benchmarking pipelines do not reliably identify models that preserve connectivity-based alignment patterns. Although RSA performs somewhat better than LP, both metrics can mis-rank models when judged against this stricter criterion of brain alignment.

5 DISCUSSION

Our main finding is that state-of-the-art vision models—despite their scale and performance—remain only weakly aligned to the human brain when evaluated with stricter, biologically grounded criteria. Conversely, brain–to–brain alignment across subjects remains consistently high. This suggests that conventional benchmarks overestimate how well models capture neural organization.

Complementary perspectives of LP and RSA. Consistent with recent work (e.g. Bo et al. (2024)), our results highlight systematic differences between linear predictivity (LP) and representational similarity analysis (RSA) for state-of-the-art vision models trained on large-scale internet image and video datasets. LP rewards high-dimensional feature spaces that can be linearly reweighted to predict neural responses, making it effective for identifying candidate "digital twins" (i.e., predictive models) of neural data. Yet this flexibility reduces its ability to discriminate among well-performing models. RSA, by contrast, compares the relational geometry of representations, which makes it more sensitive to region-specific preferences and better at distinguishing among high-scoring models. However, because RSA relies on second-order similarity, it can assign similar scores to models with different training objectives or modalities if their representational structures align. Similar to Conwell et al. (2024), we conclude that, while different metrics are complementary, they might be too flexible if the goal is to identify meaningfully brain-aligned models.

What is meaningful alignment? Meaningful alignment should reflect biologically grounded organization rather than predictive flexibility (e.g. see Nonaka et al. (2021)). As a step towards this goal, we introduced *Alignment Pattern Similarity*, which evaluates whether models capture the relational structure between brain regions (based on anatomy and/or function). Developing and standardizing such criteria remains an open challenge.

Relation to prior benchmarks Our results also refine conclusions from prior benchmarks. We replicate the finding by Sartzetaki et al. (2024) that modeling temporal dynamics is key for RSA-alignment to early visual regions, whereas models trained with semantic objectives are more aligned to higher-level regions. Tang et al. (2025) found that a single predictive objective generalized across cortical areas under LP. Using both LP and RSA, we likewise identify the same best-performing model overall. However, our results suggest that rather than reflecting a single unifying objective, this apparent generalization may instead arise from the flexibility of large feature spaces. In particular, distinct subspaces within a model's representation may be selectively exploited by linear readouts, each supporting different tasks across cortical areas. A closer analysis of these subspaces could clarify whether cross-regional alignment reflects genuine commonalities or simply the representational versatility of large models. Our results also refine conclusions from prior benchmarks. Sartzetaki & Groen (2025) reported stream-specific alignment of modular architectures, which we replicate and show to be clearer under RSA than LP. Tang et al. (2025) found that a single predictive objective generalized across cortical areas under LP. We also observe broad alignment with LP, but suggest this may reflect the flexibility of large feature spaces rather than a unified computational principle. Finally, large-scale efforts such as BrainScore (Schrimpf et al., 2018) and the Algonauts challenges (Gifford et al., 2024; 2023; Cichy et al., 2021) have advanced the field, but their reliance on LP may systematically overstate alignment. We extend these frameworks in two directions: (i) by analyzing the number of *practically equivalent* models and (ii) by introducing APS as an anatomically grounded measure. In both analyses, RSA proved more informative than LP.

Implications We find that current normative models—even if brain-aligned in the sense of standard metrics like RSA and LP—do not faithfully reproduce brain-to-brain alignment patterns when substituted for real neural data. Progress will require the field to define and agree on stricter alignment criteria to complement existing measures. The key challenge is not only to build more powerful models, but also to evaluate them against criteria that more directly capture the organization of the brain.

REFERENCES

Dimitris Achlioptas. Database-friendly random projections. *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2001. URL `https:`

//api.semanticscholar.org/CorpusID:2640788.

Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003. ISSN 0022-0000. doi: https://doi.org/10.1016/S0022-0000(03)00025-4. URL https://www.sciencedirect.com/science/article/pii/S0022000003000254. Special Issue on PODS 2001.

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.*, 25(1):116–126, January 2022.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025a. URL https://arxiv.org/abs/2506.09985.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025b.

Daniel M Bear, Kevin Feigelis, Honglin Chen, Wanhee Lee, Rahul Venkatesh, Klemen Kotar, Alex Durango, and Daniel L K Yamins. Unifying (machine) vision via counterfactual world modeling. *arXiv [cs.CV]*, June 2023.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *CoRR*, abs/2102.05095, 2021. URL https://arxiv.org/abs/2102.05095.

Yiqing Bo, Ansh Soni, Sudhanshu Srivastava, and Meenakshi Khosla. Evaluating representational similarity measures from the lens of functional correspondence. *arXiv preprint arXiv:2411.14633*, 2024.

Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E Hummel, Rachel F Heaton, Benjamin D Evans, Jeffrey Mitchell, and Ryan Blything. Deep problems with neural network models of human vision. *Behav. Brain Sci.*, 46:e385, December 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. URL https://arxiv.org/abs/2002.05709.

R M Cichy, K Dwivedi, B Lahner, A Lascelles, P Iamshchinina, M Graumann, A Andonian, N A R Murty, K Kay, G Roig, and A Oliva. The algonauts project 2021 challenge: How the human brain makes sense of a world in motion. *arXiv [cs.CV]*, April 2021.

MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020.

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. *Nat. Commun.*, 15(1):9383, October 2024.

Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. TRIBE: TRImodal brain encoder for whole-brain fMRI response prediction. *arXiv [cs.LG]*, July 2025.

A T Gifford, B Lahner, S Saba-Sadiya, M G Vilas, A Lascelles, A Oliva, K Kay, G Roig, and R M Cichy. The algonauts project 2023 challenge: How the human brain makes sense of natural scenes. *arXiv [cs.CV]*, January 2023.

Alessandro T Gifford, Domenic Bersch, Marie St-Laurent, Basile Pinsard, Julie Boyle, Lune Bellec, Aude Oliva, Gemma Roig, and Radoslaw M Cichy. The algonauts project 2025 challenge: How the human brain makes sense of multimodal movies. *arXiv [q-bio.NC]*, December 2024.

Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

M A Goodale and A D Milner. Separate visual pathways for perception and action. *Trends Neurosci.*, 15(1):20–25, January 1992.

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. https://github.com/facebookresearch/vissl, 2021.

Emily D Grossman and Randolph Blake. Brain areas active during visual perception of biological motion. *Neuron*, 35(6):1167–1175, 2002.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Dwight J Kravitz, Kadharbatcha S Saleem, Chris I Baker, and Mortimer Mishkin. A new neural framework for visuospatial processing. *Nat. Rev. Neurosci.*, 12(4):217–230, April 2011.

Nikolaus Kriegeskorte and Xue Xin Wei. Neural tuning and representational geometry. *Nature Reviews Neuroscience*, 22:703–718, 2021. ISSN 14710048. doi: 10.1038/s41583-021-00502-3. URL http://dx.doi.org/10.1038/s41583-021-00502-3.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. 2, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008. URL https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008/full. Publisher: Frontiers.

Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N. Apurva Ratan Murty, Kendrick Kay, Aude Oliva, and Radoslaw Cichy. Modeling short visual events through the BOLD moments video fMRI dataset and metadata. 15(1):6241, 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-50310-3. URL https://www.nature.com/articles/s41467-024-50310-3. Publisher: Nature Publishing Group.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3202–3211, 2022a.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022b. URL https://arxiv.org/abs/2201.03545.

Soma Nonaka, Kei Majima, Shuntaro C. Aoki, and Yukiyasu Kamitani. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*, 24, 9 2021. ISSN 25890042. doi: 10.1016/j.isci.2021.103013.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Edmund T Rolls. Two what, two where, visual cortical streams in humans. *Neurosci. Biobehav. Rev.*, 160:105650, May 2024.

Christina Sartzetaki and Iris I A Groen. Mapping modular processing of compressed videos across human visual cortex. August 2025.

Christina Sartzetaki, Gemma Roig, Cees GM Snoek, and Iris IA Groen. One hundred neural networks and brains watching videos: Lessons from alignment. *bioRxiv*, pp. 2024–12, 2024.

Alexander Sax, Bradley Emi, Amir R. Zamir, Leonidas J. Guibas, Silvio Savarese, and Jitendra Malik. Mid-level visual representations improve generalization and sample efficiency for learning visuomotor policies. 2018.

Rylan Schaeffer, Mikail Khona, Sarthak Chandra, Mitchell Ostrow, Brando Miranda, and Sanmi Koyejo. Position: Maximizing neural regression scores may not identify good models of the brain. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, October 2024.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2018.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Ansh Soni, Sudhanshu Srivastava, Konrad Kording, and Meenakshi Khosla. Conclusions about neural network to brain alignment are profoundly impacted by the similarity measure. *bioRxiv*, pp. 2024.08.07.607035, August 2024.

Stefan Stojanov, David Wendt, Seungwoo Kim, Rahul Venkatesh, Kevin Feigelis, Jiajun Wu, and Daniel LK Yamins. Self-supervised learning of motion concepts by optimizing counterfactuals, 2025. URL https://arxiv.org/abs/2503.19953.

Stefan Sunaert, Paul Van Hecke, Guy Marchal, and Guy A Orban. Motion-responsive regions of the human brain. *Experimental brain research*, 127(4):355–370, 1999.

Yingtian Tang, Abdulkadir Gokce, Khaled Jedoui Al-Karkari, Daniel Yamins, and Martin Schrimpf. Many-two-one: Diverse representations across visual pathways emerge from a single objective. *bioRxiv*, pp. 2025.07.22.664908, July 2025.

Alexander Walther, Hamed Nili, Naveed Ejaz, Arjen Alink, Nikolaus Kriegeskorte, and Jörn Diedrichsen. Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage*, 137:188–200, 2016.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Jialin Wu, Shreya Saha, Yiqing Bo, and Meenakshi Khosla. Measuring the measures: Discriminative capacity of representational similarity metrics across model families. *arXiv [cs.LG]*, September 2025.

Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. 19(3):356–365, March 2016.

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.

Amir Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning, 2018. URL `https://arxiv.org/abs/1804.08328`.

A Appendix

A.1 Disclosure of LLM use

We have used LLMs to assist in the code writing process, including for plot creation, to discuss ideas and concepts, in literature search, for searching information in a given work, and for refining text in this paper.

A.2 Noise Ceilings

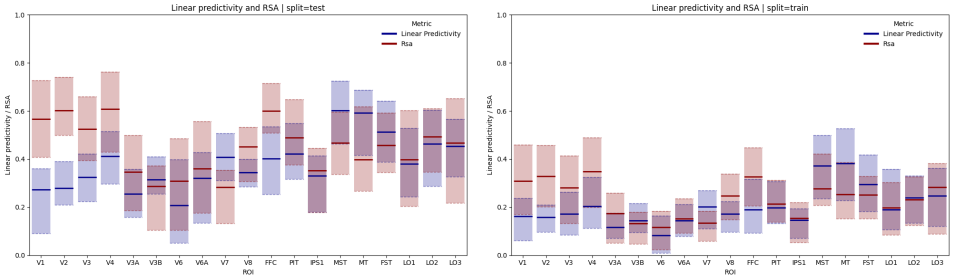

Figure 6: Noise ceilings per ROI. Left: test split, right: train split of BOLDMoments. For each ROI: Middle line: mean across all subject, upper and lower line: highest and lowest noise ceiling across subjects.

A.3 Model Alignment by Visual Areas

A.4 Equivalent Models by Visual Area

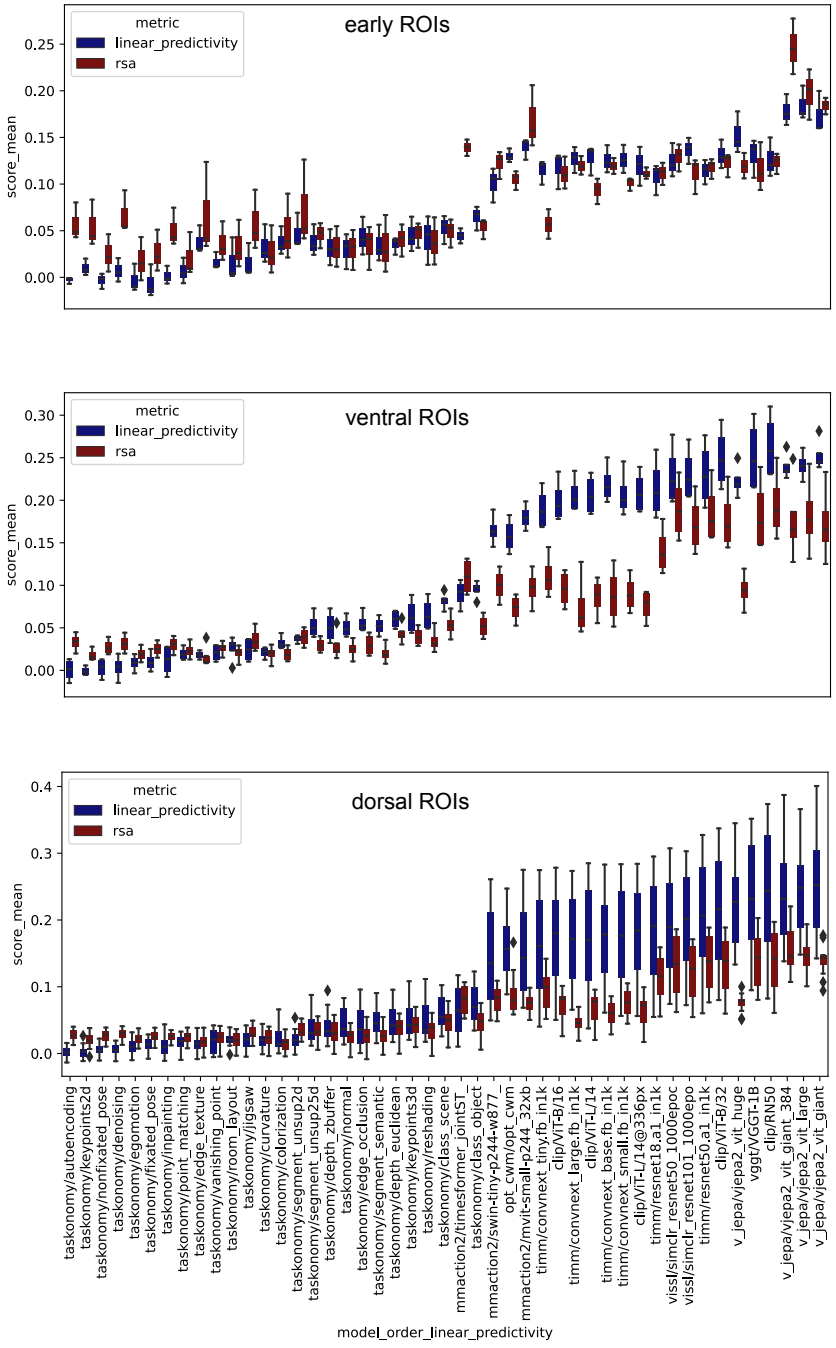

Figure 7: Alignment scores of all models in early layers, dorsal stream, and ventral stream. Boxplots show scores across ROIs.

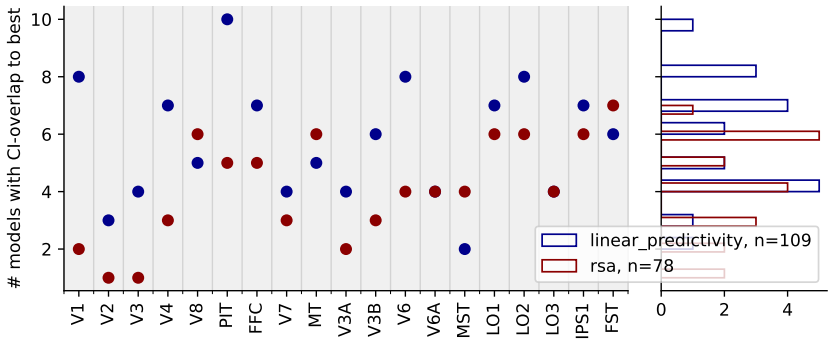

Figure 8: Number of equivalent models for each ROI and metric.

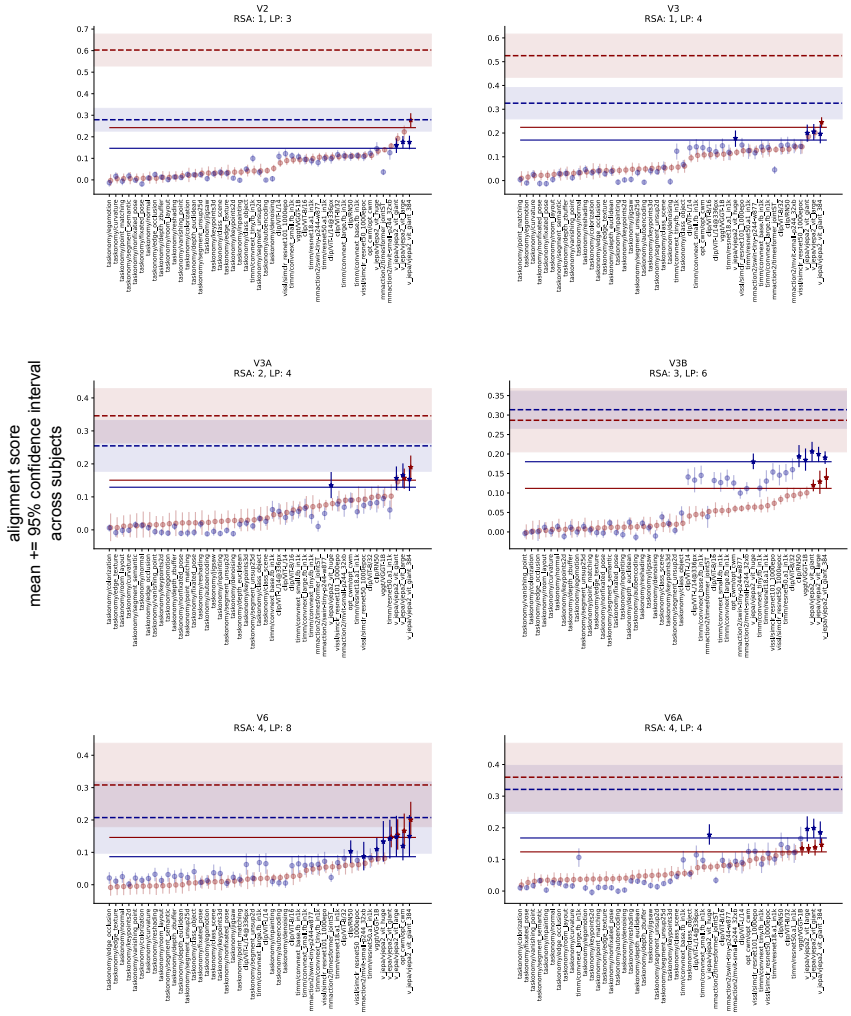

Figure 9: Model rankings for individual ROIs.

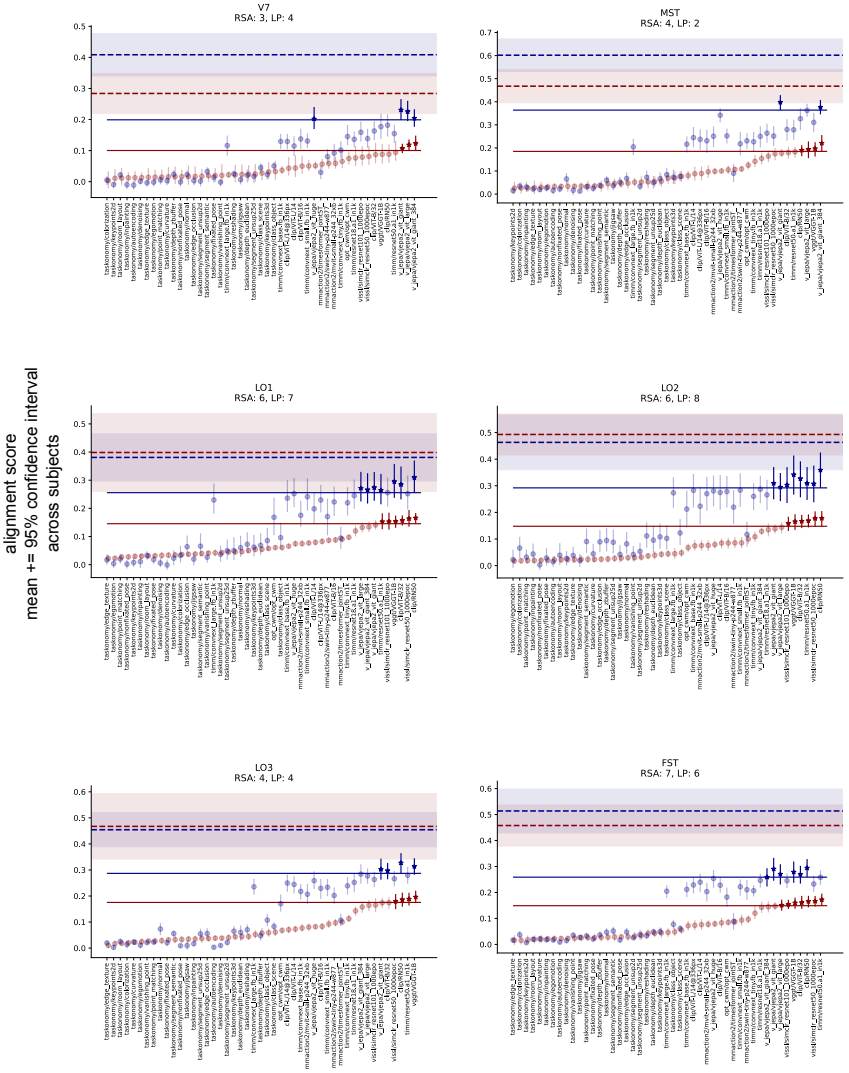

Figure 10: Model rankings for individual ROIs.

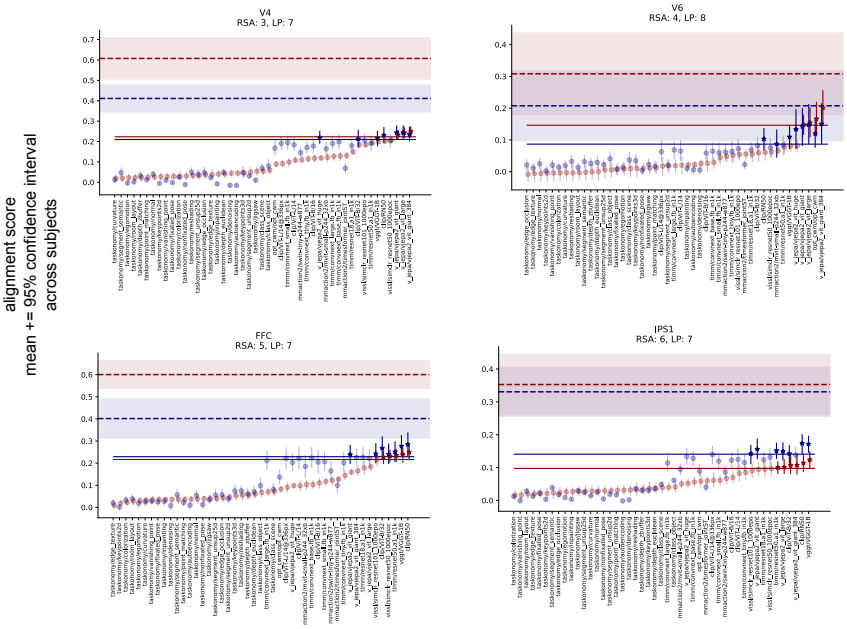

Figure 11: Model rankings for individual ROIs.