Only Brains Align with Brains: Cross-Region Patterns Expose Limits of Normative Models

Anonymous authors
Paper under double-blind review

Abstract

Neuroscientists and computer vision scientists alike have relied on model-brain alignment benchmarks to find parallels between artificial and biological vision systems. These benchmarks rank models according to alignment measures (AM) such as representational similarity analysis (RSA) and linear predictivity (LP). However, recent works have revealed a number of problems with these rankings, such as their sensitivity towards the choice of AM, raising the deeper conceptual question of what it means for a model to be "brain-aligned." Here, we introduce the notion of *alignment patterns* - characteristic patterns of alignment between brain regions - and posit that models should reproduce these patterns in order to be considered brain-aligned. First, we apply a standard benchmarking pipeline to a broad spectrum of vision models on the BOLD-Moments video fMRI dataset across visual regions of interest (ROIs). We find that, while this pipeline can identify nominally best predictive models, many other models fall within subject-level variability and are therefore practically equivalent in terms of brain alignment. We then apply our complementary relational criterion: a ROI-aligned model should reproduce that ROIs cross-region alignment pattern. We find that, while these patterns are highly stable across brains of different subjects, even top-ranked models fail to capture them. Notably, models that appear practically equivalent in predictive accuracy diverge sharply under the relational criterion, revealing both the limitations with respect to discriminative power of existing evaluation pipelines, as well as alignment pattern analysis as a way of increasing this discriminative power. Finally, we argue for a principled distinction between brain-predictivity and brain-alignment. For applications such as digital twins, prediction performance may suffice; but for understanding the inductive biases of the visual system, models should satisfy stricter distributional and relational criteria.

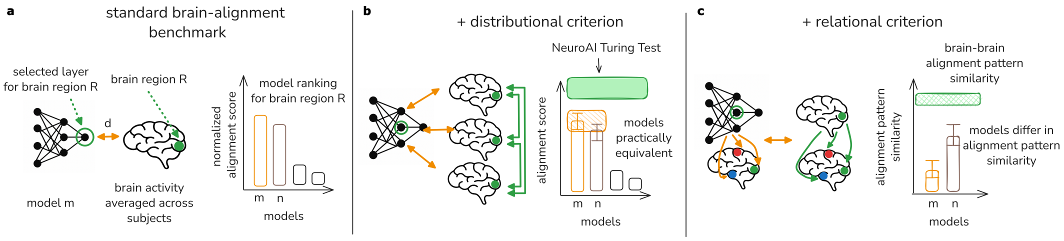
a
standard brain-alignment benchmark
selected layer for brain region R
brain region R
d
model ranking for brain region R
normalized alignment score
model m
brain activity averaged across subjects
m
n
models
b
+ distributional criterion
NeuroAI Turing Test
alignment score
models practically equivalent
m
n
models
c
+ relational criterion
brain-brain alignment pattern similarity
alignment pattern similarity
models differ in alignment pattern similarity
m
n
models

Figure 1: **Extending brain-alignment benchmarks with distributional and relational criteria.** **(a)** Standard brain-alignment benchmarks report model rankings based on normalized and averaged scores obtained from some alignment measure d, obscuring both the gap to ceiling performance, as well as the degree of variability in predictivity across the population. **(b** In addition to assessing the distribution of brain-brain alignment scores across the population ("NeuroAI Turing Test")(Feather et al., 2025; Thobani et al., 2025), we assess the distribution of model-brain alignment scores to define equivalence classes of models in terms of brain-alignment. **(c)** To distinguish between equivalently aligned models, we introduce alignment pattern similarity as a relational criterion that assesses whether a model reproduces the cross-region alignment patterns of visual regions of interest (ROIs).

1 Introduction

A central aim of vision research is to understand which evolutionary and developmental inductive biases have shaped the biological - and, in particular, 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 Yamins et al. (2014); Yamins & DiCarlo (2016). In this line of research, models trained on a variety of tasks from computer vision are evaluated with regard to how well they *align* to functional recordings from visual brain areas according to a variety of measures (Klabunde et al., 2025), and high alignment is typically interpreted as evidence that a model's task, architecture, or training dataset reflect biologically relevant constraints. Recent progress in computer vision, including large-scale self-supervised and multi-modal models (e.g. Bear et al., 2023; Assran et al., 2025), along with the systematic collection of large-scale fMRI datasets (Lahner et al., 2024; Allen et al., 2022) now allow to scale these efforts towards comprehensive brain-alignment benchmarks of vision models (Schrimpf et al., 2018; 2020; Conwell et al., 2024; Sartzetaki et al., 2024). The conclusions drawn from such benchmarks, however, rest on the assumption that alignment score rankings reflect differences in brain–alignment in some meaningful way. Recent work has challenged this assumption, showing that commonly used measures yield inconsistent results, and thus 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"?* Previous works (Feather et al., 2025; Thobani et al., 2025) have posited that a model should only be considered truly brain-like if its internal representations are indistinguishable - under a similarity transform - from those of other brains. The authors suggest a *distributional* criterion to assess this: if the model-brain alignment score is within the distribution of brain-brain alignment scores, the model is considered indistinguishable from other brains under the similarity transform that generated the score, thereby passing the *NeuroAI Turing Test* (Fig. 1 a). The authors apply their distributional criterion to a range of static, i.e. image-based model-brain alignment benchmarks and conclude that many of them are saturated, i.e., models pass the NeuroAI Turing Test (Feather et al., 2025). Here, we make three major contributions on top of existing benchmarks and the NeuroAI Turing Test:

1. **Benchmarking**
   First, we evaluate a broad family of state-of-the-art vision models on a naturalistic video fMRI dataset spanning the entire visual hierarchy (Lahner et al., 2024), resolving brain-alignment to individual visual ROIs (Fig. 1a). We apply the NeuroAI Turing test and find that many models pass the NeuroAI Turing test when evaluated with Linear Predictivity (LP), but most fail when evaluated with Representational Similarity Analysis (RSA). This highlights the *lack of robustness* of standard benchmarking pipelines to the choice of alignment measure.
2. **Defining practical equivalence in brain-alignment**
   Second, using a distributional criterion to analyze alignment-score rankings, we show that for most visual ROIs, models that differ substantially in architecture, training data, and objective are still *practically equivalent* in their brain-alignment (Fig. 1b). This challenges the discriminative power of current benchmarks.
3. **Alignment pattern similarity to distinguish among equivalently aligned models**
   Third, we introduce a *relational* criterion to distinguish among equivalently aligned models: We posit that a model should only be considered aligned to a brain region if it reproduces the *cross-region alignment pattern* of this brain region (Fig. 1c).

We conclude with a discussion where we argue to make a distinction between *brain-aligned* models in this stricter sense, and *brain-predictive* models, where high LP-scores in suffice.

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.

Alignment patterns. A number of works have evaluated whether the visual hierarchy of regions or voxels is reflected in the order of their best-matching layers within a DNN, e.g., Güçlü & Van Gerven (2015); Cichy et al. (2016); Bersch et al. (2025); Thobani et al. (2025). Nonaka et al. (2021) suggest using such hierarchical correspondence as an alignment criterion. This evaluates whether entire models match to entire visual streams, while we look at alignment of models to individual regions. Thobani et al. (2025) also correlate layer dissimilarity scores to known distances between layers, and repeat the same analysis for visual brain regions, which they assign to integer hierarchy levels 1 to 5. This serves to compare alignment methods, not to evaluate model-brain similarity.

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 1000 stimuli in the train split was shown three times to each subject, each of the 102 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, dropping data for 2 (4) stimuli from the train (test) set because of frame extraction issues, 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

Because fMRI data are noisy, neither perfect predictivity nor perfect representational similarity can be expected (Walther et al., 2016). We compute two noise ceilings for each ROI in the following ways: **Upper noise ceiling.** We average the fMRI data of N-1 subjects for the given ROI. Then we use this average as predictor feature space and compute RSA/LP score between the average map and the remaining subject's ROI data. This yields one upper noise ceiling per subject, from which we can compute a mean, and a 95% confidence interval of the mean, across the ten subjects. **Lower noise ceiling.** As suggested in the Neuro-AI Turing test Feather et al. (2025), we compute a noise ceiling based on pairwise alignment scores between subjects. For a given ROI, we sample five other subjects per target subject, excluding previously sampled pairs, for a total of 50 subject pairs. For each pair we compute the RSA/LP score between the regions' fMRI features of the two subjects, and again compute mean and 95% confidence interval across all pairs.

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., 2025) 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 block 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 ($1 -$ Pearson correlation) for all samples. The overall RSA alignment score is then the Pearson 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 determine when models are practically indistinguishable in terms of brain alignment, we defined a practical equivalence criterion based on bootstrap estimates of variability in model-brain alignment scores. For a given model m with feature space X_m, we generated a bootstrap distribution of mean brain-alignment scores under a measure $\mathcal{M}$ by resampling subject indices with replacement. Specifically, we defined a bootstrap index vector

$$I^* = (i_1^*, \dots, i_{10}^*), \qquad i_k^* \sim \text{Unif}(I) \text{ with replacement}, \quad I = \{1, \dots, 10\},$$

and computed the corresponding bootstrap estimate of the mean alignment score as

$$\frac{1}{10} \sum_{k=1}^{10} \mathcal{M}\big(X_m, Y_{i_k^*}\big).$$

We then derived 95% confidence intervals for the model's mean brain-alignment score from the resulting distribution. A model m was deemed practically equivalent to the top-ranking model t if its empirical mean alignment score $\langle \mathcal{M}(X_m, Y_i) \rangle_i$ fell within the 95% confidence interval of the top ranking model.

3.5 Alignment Pattern Analysis

We define an alignment pattern α under a similarity transform $\mathcal{M}$ between a predictor feature space ϕ_p and N target feature spaces $\Psi_t = [\psi_t^1, ..., \psi_t^N]$ as

$$\alpha(\phi_p, \Psi_t) = [\mathcal{M}(\phi_p, \psi_t^1), \mathcal{M}(\phi_p, \psi_t^2), ..., \mathcal{M}(\phi_p, \psi_t^N)] \tag{1}$$

3.5.1 fMRI-derived alignment patterns

For fMRI-derived alignment patterns, both the predictor and the target feature spaces are sourced from brain activity from the BOLDMoments dataset. fMRI-derived alignment patterns are defined between pairs of subjects p, t, where the brain activity of subject p functions as the predictor feature space ϕ_p and the brain activity of subject t functions as the target feature space Ψ_t The alignment pattern for a given ROI $r \in N$ and a pair of subjects p, t is then defined as

$$\alpha_r(\phi_p, \Psi_t) = [\mathcal{M}(\phi_p^r, \psi_t^1), \mathcal{M}(\phi_p^r, \psi_t^2), ..., \mathcal{M}(\phi_p^r, \psi_t^r), ..., \mathcal{M}(\phi_p^r, \psi_t^N] \tag{2}$$

We detail in the Appendix Section S3.1 how the variance of fMRI-derived alignment patterns is estimated.

3.5.2 Model-derived alignment patterns

For model-derived alignment patterns, the predictor feature space is defined as the activations in one layer l of the model, ϕ_m^l, and the target feature spaces are analogous to the case of fMRI-derived alignment patterns. A model-derived alignment pattern between model m and subject t for a given ROI $r \in N$ is then

$$\alpha_l(\phi_m, \Psi_t) = [\mathcal{M}(\phi_m^l, \psi_t^1), \mathcal{M}(\phi_m^l, \psi_t^2), ..., \mathcal{M}(\phi_m^l, \psi_t^N] \tag{3}$$

3.5.3 Structural connectivity-derived alignment patterns

For comparing alignment patterns to structural connectivity patterns, we use a network based on diffusion-weighted tensor imaging (DTI) Pierpaoli et al. (1996) streamline-density from the Human Connectome Young Adult full dataset (Caron & Pestilli, 2023) as provided through brainlife Hayashi et al. (2024) (provided as 'conmat' datatype). The procedure fits streamlines - white-matter trajectory candidates Smith et al. (2012) - to diffusion MRI data. The number of streamlines intersecting both ROIs of a pair of regions is divided by the volume of both regions to obtain the 'density'-based connectivity matrix we use. For more information, see Hayashi et al. (2024), section *dMRI processing*. We average the connectivity matrices of 1065 subjects to obtain a single connectivity matrix, $C = (c_{r,t})_{r,t=1\cdots N}$ where $c_{r,t}$ is the streamline density between regions r and t. The structural connectivity-derived alignment pattern for a given ROI r is then

$$\alpha_{struct}(r) = [c_{r,1}\ ,\ \dots\ ,\ c_{r,r-1}\ ,\ c_{r,r+1}\ ,\ \dots\ ,\ c_{r,N}] \tag{4}$$

where we exclude the ROI r since self-similarity is not defined for streamline-density as alignment measure.

3.5.4 Alignment pattern similarity

Alignment pattern similarity between two alignment patterns, e.g. a fMRI-derived alignment pattern $\alpha_r(\phi_p, \Psi_t)$ and a model-derived alignment pattern $\alpha_l(\phi_m, \Psi_t)$ is calculated as

$$\rho(\alpha_r(\phi_p, \Psi_t), \alpha_l(\phi_m, \Psi_t)) \tag{5}$$

where ρ is Pearson's correlation coefficient.

4 Results

4.1 Benchmarking alignment of vision models to the visual cortex

4.1.1 V-JEPA achieves highest overall brain-alignment scores

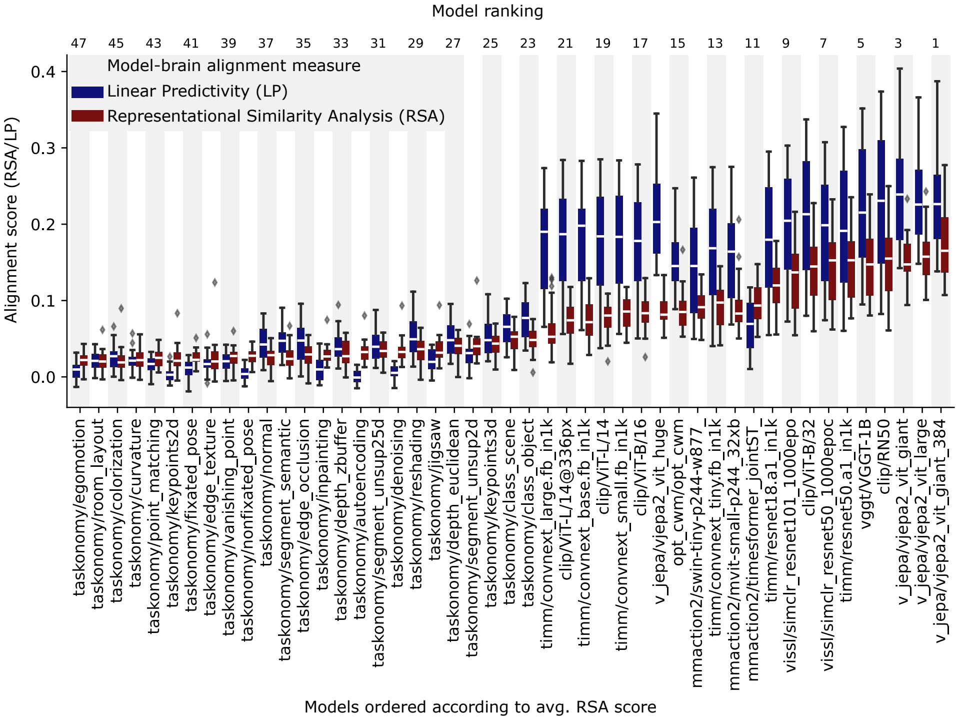

Figure 2: **Standard benchmarking results for the BOLDMoments dataset**. Boxplots depict the distribution of subject-averaged alignment scores (RSA/LP) across ROIs.

We evaluated a broad range of vision models with respect to their alignment to visual cortex—including early, ventral, and dorsal regions—using two complementary alignment measures: RSA and LP. The models varied with respect to architecture (CNNs and Transformers), training objective

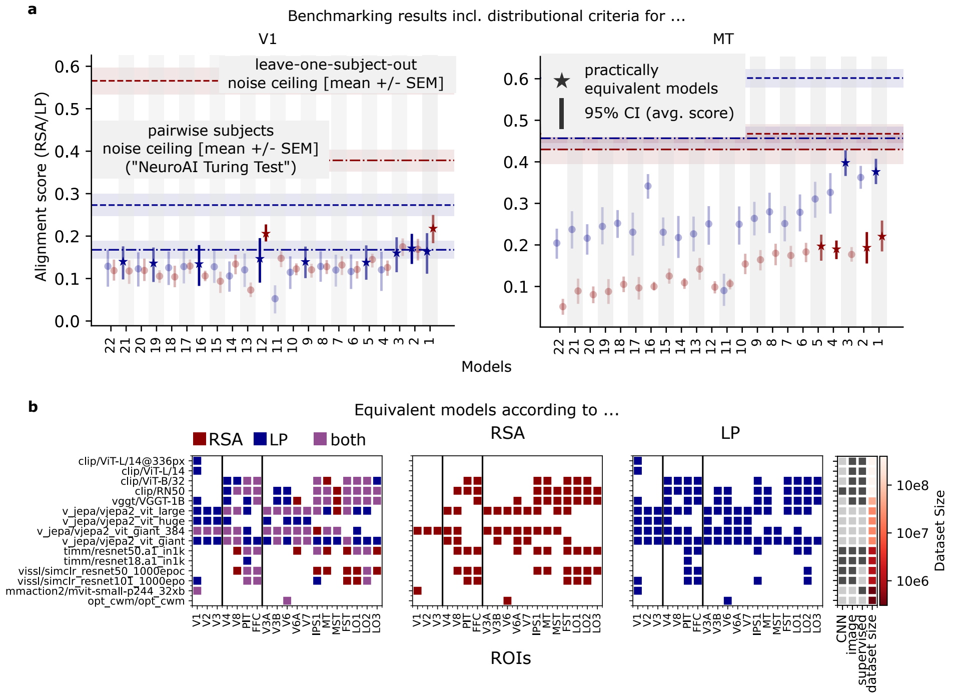

Figure 3: **ROI-wise benchmarking results including noise ceilings and practical equivalence assessment a** Benchmarking results for the top 22 models for two example regions, including upper and lower noise ceilings (see Sec.3.1.1), as well as 95% CIs about the mean alignment score for each model. Practically equivalent models (see Sec. 3.4 are indicated by stars. **b** Overview of all models that are practically equivalent for a given ROI according to RSA, LP or both. The right-most box shows properties of the models.

(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., 2025), achieved the strongest overall alignment scores across visual cortex, according to both RSA and LP (Fig. 2). Notably, however, the best aligned models included CLIP with a ResNet-50 backbone and the VGG-Transformer—which differ in several important aspects from V-JEPA 2 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.

4.1.2 Models pass the NeuroAI Turing test on LP, but not on RSA

We compared model-brain alignment scores to an upper and a lower noise ceiling derived from inter-subject (i.e., brain-brain) alignment distributions. The lower noise ceiling is defined as the 95% confidence interval (CI) of the average alignment of brain activity of any two subjects in the population(Feather et al., 2025; Thobani et al., 2025) ("NeuroAI Turing Test"). The upper noise ceiling is defined as the 95% confidence interval (CI) of the average alignment of the leave-one-subject-out average brain activity to the left-out subject. We find that for LP, many models reach or even surpass the lower noise ceiling for many ROIs, thereby passing the NeuroAI Turing test. However, both the noise ceiling and the absolute model performance are relatively low. In contrast, for RSA this is rarely the case - namely, for the ROI-model combinations V-JEPA ViT-giant-384 - V3A, V-JEPA ViT-giant-384 - V6, V-JEPA ViT-large - V6, and Opt-CWM - V6 (Fig. 3a, Figs. S4.1–S4.7).

4.2 Different models are practically equivalent in model-brain alignment

Next, we further assessed the robustness of model rankings and the discriminability between models in terms of brain alignment by checking for models that were practically equivalent in brain-alignment (see Methods 3.4), i.e., whose scores fell within the 95% CI of the mean score of the top-ranking model. This analysis revealed that many models were practically equivalent in terms of brain alignment (Fig. 3 a, b; Figs. S4.1 - S4.7), but results differed between LP and RSA. For example, in primary visual cortex (V1), LP grouped nine models as practically equivalent to the best model, whereas RSA reduced this set to just two. We found a similar pattern 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 4.1 models as practically equivalent per region, compared to 5.6 with LP.

These findings demonstrate the lack of discriminative power of alignment-measure based model rankings, motivating the need for additional criteria to distinguish between equivalently aligned models.

4.3 Alignment pattern similarity as a necessary criterion for brain-alignment

We propose a necessary (though not sufficient; see Discussion) criterion for alignment to a brain region: a model should not only match that region locally, but also preserve its pattern of relationships to other regions. First, we estimated fMRI-derived cross-region alignment patterns. These patterns are highly consistent *within* each ROI for both RSA and LP (RSA: Fig. 4a,b, black lines and boxes; LP: Suppl. Fig. S4.9). Moreover, whereas RSA yields clearly distinct alignment patterns for different ROIs, LP does not: LP-based cross-region patterns are substantially more homogeneous across ROIs (Suppl. Fig. S4.9).

Next, we examined *model*-derived alignment patterns for all models in the equivalence class of each ROI (Section 4.2). RSA produces strongly model-specific alignment patterns, enabling discrimination among models that are otherwise equivalent (Fig. 4a,b, colored lines and boxes). In contrast, LP yields uniformly high pattern similarity across models, providing little discriminability (Suppl. Fig. S4.9). Training further increases alignment-pattern similarity (APS) for RSA—often substantially—but has little to no effect on LP-based APS (Suppl. Fig. S4.10).

Finally, applying a lenient criterion that a model's alignment-pattern similarity to its ROI's pattern must at least be positive, RSA-based APS excludes three V-JEPA variants as candidate models across ten (mostly dorsal) ROIs (Fig. 4c). In contrast, LP-based APS excludes only three out of nine candidate models, and only for a single ROI (V1).

Overall, these results show that RSA-based fMRI-derived APs are highly reliable and ROI-specific, whereas LP-derived patterns are both more homogeneous across ROIs and similarly high for nearly all models, trained or untrained. Consequently, LP offers limited ability to distinguish between models on the basis of cross-region alignment structure.

4.4 RSA-based but not LP-based alignment patterns track structural connectivity

To better understand the factors determining brain-brain alignment patterns, and the role of the alignment measure used to calculate APs, we estimated APs from an independent dataset of structural connectivity from N=1065 humans(Caron & Pestilli, 2023). The similarity measure used to calculate connectivity-derived APs is streamline density(Pierpaoli et al., 1996), which takes the role of RSA/LP in the fMRI-based APs. We then compared these connectivity-derived APs with RSA-based and LP-based fMRI-derived APs. We found that RSA-based fMRI-derived APs were similar to connectivity-derived APs for most early, ventral and late dorsal regions (Fig. 5, 12/19 ROIs significant; Methods), whereas the same analysis for LP-based APs revealed much lower similarity between fMRI-derived and connectivity-derived APs (Suppl. Fig. S4.11, 5/19 ROIs significant).

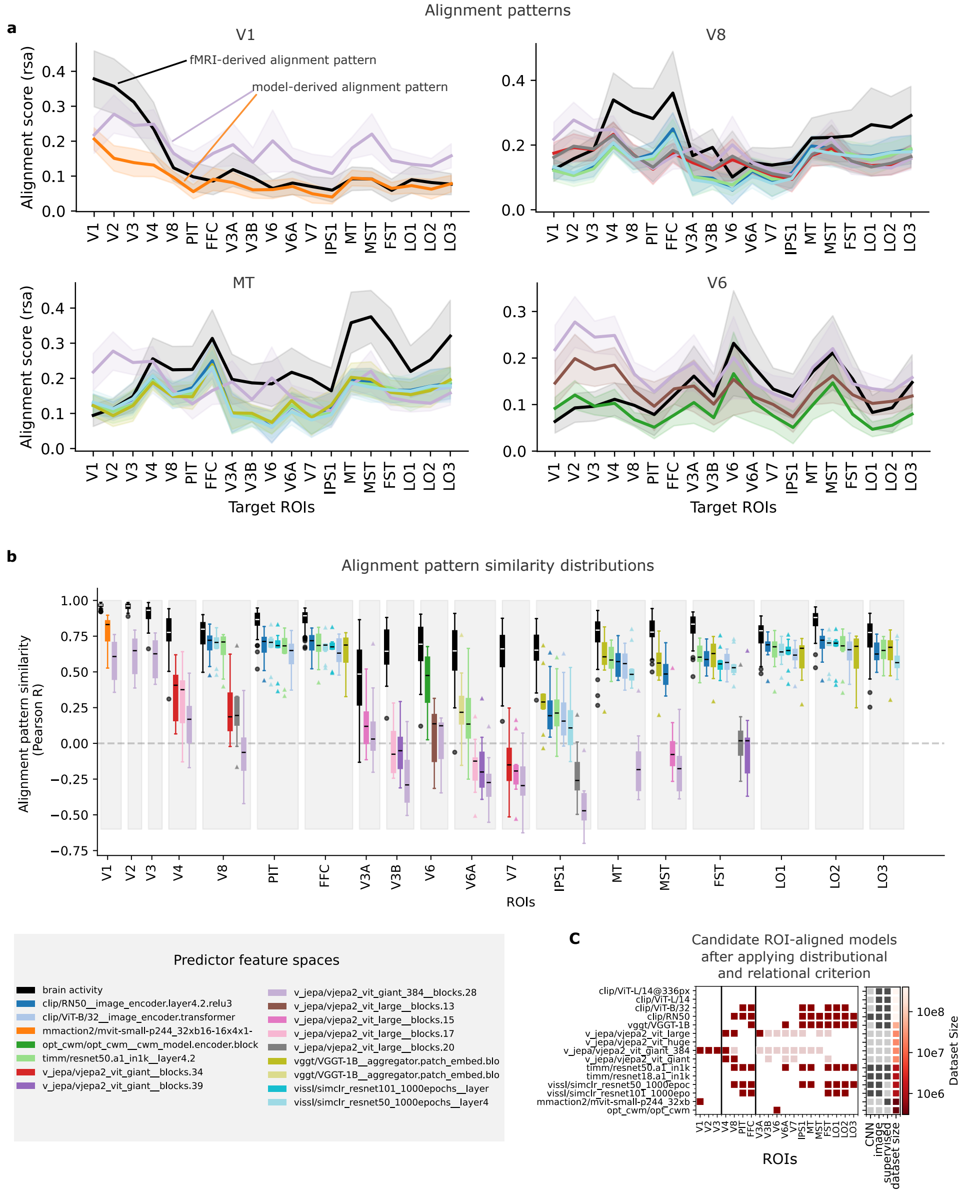

Figure 4: Alignment patterns are consistent within ROIs and distinguish equivalently aligned models. **(a)** RSA-based fMRI-derived (black) and model-derived (coloured) alignment patterns for four example ROIs. Shaded area indicates +-SD (see Appendix Sec. S3.1). **(b)** Distributions of alignment pattern similarities within a ROI (black box-plots) and between a ROI and its equivalently aligned models (coloured box-plots). **(c)** Same plot as Fig. 3b, but graying out models whose 95% CI of the mean APS includes zero.

5 Discussion

In this work, we show that model rankings produced by standard brain-alignment benchmarking pipelines are insufficient both for identifying the most brain-aligned models as well as for distinguishing them from merely brain-predictive models. To alleviate this problem, we introduce and apply a relational criterion - alignment pattern similarity - and propose to use it as an additional criterion in alignment benchmarks to increase their discriminative power.

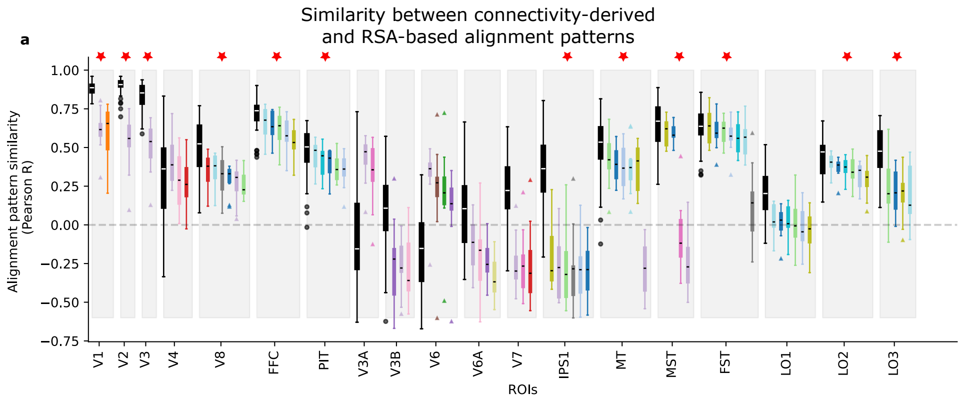

Figure 5: Connectivity-derived alignment pattern similarity across ROIs and models. Colour code as in previous plots. APS is calculated between RSA-based fMRI-derived and connectivity-derived APs (black box-plots), and between RSA-based model-derived and connectivity-derived APs (coloured box-plots) (see Methods Sec. 3.5.2). Stars indicate significantly higher fMRI-connectivity APS than with random connectivity-derived APs (see Appendix Sec. S3.2).

What does it mean to be "brain-aligned"? As pointed out by Schaeffer et al. (Schaeffer et al., 2024), "NeuroAI lacks canonical definitions of neural similarity". This lack of canonical definitions, among other factors, underlies recent discussions about common practices of brain-alignment benchmarking and the kinds of conclusions that can (or cannot) be drawn about model-brain similarity based on the results of such benchmarks (e.g. Dujmovi et al. (2024); Bowers et al. (2023)). At the core of the discussion is the repeated finding that models often achieve high brain-alignment according to some measure while diverging from the brain in other aspects that neuroscientists consider relevant to "true" brain-alignment (Schaeffer et al., 2022; Malhotra & Bowers, 2024), without the field agreeing on what those aspects are. Here, we propose alignment pattern similarity as an additional *necessary, but not sufficient criterion* for brain-alignment: low APS excludes models from the pool of potentially brain-aligned models, but high APS does not confirm brain-alignment of a model.

Alignment patterns reveal additional implicit biases of alignment measures Consistent with recent work (e.g. Soni 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. It has been conjectured that LP implicitly rewards higher-dimensional predictor feature spaces (Schaeffer et al., 2024), which warrants caution in drawing conclusions about brain-alignment from LP-based model rankings. Our results are consistent with this conjecture: LP-based rankings mostly separate models with richer feature spaces from those with less rich feature spaces. Here, in addition to this, we find that LP-based alignment patterns are similar in shape (Suppl. Fig.S4.8), leading to a second conjecture: LP scores carry implicit biases not only about the richness of the predictor feature spaces, but also about the *predictability of the target feature spaces*.

Distinguishing between brain-predictivity and brain-alignment These findings motivate a clear distinction: LP may be effective at identifying brain-predictive models - which can be very useful e.g. as digital twins in a variety of settings such as BMI-applications - but insufficient for discriminating between more and less brain-aligned models in a stricter sense of the term. Recent works have made progress towards narrowing down the set of candidate brain-aligned models by increasing the discriminative power of benchmarks. A promising approach relies on combining complementary measures (Wu et al., 2025). We propose APS in a similar spirit: as a biologically motivated complementary measure for brain-alignment.

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, 2025. URL https://arxiv.org/abs/2506.09985.

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.

Domenic Bersch, Martina G Vilas, Sari Saba-Sadiya, Timothy Schaumlöffel, Kshitij Dwivedi, Christina Sartzetaki, Radoslaw M Cichy, and Gemma Roig. Net2brain: A toolbox to compare artificial vision models with human brain responses. *Frontiers in Neuroinformatics*, 19:1515873, 2025.

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.

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. *Behavioral and Brain Sciences*, 46, 12 2023. ISSN 14691825. doi: 10.1017/S0140525X22002813.

Brad Caron and Franco Pestilli. Brainlife paper - human connectome young adult - full dataset, 2023. URL https://brainlife.io/pub/640a3f9dc538c16a826f9b1a.

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.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):27755, 2016.

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.

Marin Dujmovi, Jeffrey S Bowers, Federico Adolfi, and Gaurav Malhotra. Inferring dnn-brain alignment using representational similarity analysis can be problematic. Technical report, 2024.

Jenelle Feather, Meenakshi Khosla, N. Apurva Ratan Murty, and Aran Nayebi. Brain-model evaluations need the neuroai turing test, 2025. URL `https://arxiv.org/abs/2502.16238`.

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.

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.

Umut Güçlü and Marcel AJ Van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

Soichi Hayashi, Bradley A Caron, Anibal Sólon Heinsfeld, Sophia Vinci-Booher, Brent McPherson, Daniel N Bullock, Giulia Bertò, Guiomar Niso, Sandra Hanekamp, Daniel Levitas, et al. brainlife. io: a decentralized and open-source cloud platform to support neuroscience research. *Nature methods*, 21(5):809–813, 2024.

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.

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.

Gaurav Malhotra and Jeffrey Bowers. Predicting brain activation does not license conclusions regarding dnn-brain alignment: The case of brain-score. 2024. URL http://arxiv.org/abs/1811.12231.

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.

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.

Carlo Pierpaoli, Peter Jezzard, Peter J Basser, Alan Barnett, and Giovanni Di Chiro. Diffusion tensor mr imaging of the human brain. *Radiology*, 201(3):637–648, 1996.

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.

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, and Ila Fiete. No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 16052–16067. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/66808849a9f5d8e2d00dbdc844de6333-Paper-Conference.pdf.

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, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020.

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.

Robert E Smith, Jacques-Donald Tournier, Fernando Calamante, and Alan Connelly. Anatomically-constrained tractography: improved diffusion mri streamlines tractography through effective use of anatomical information. *Neuroimage*, 62(3):1924–1938, 2012.

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.

Imran Thobani, Javier Sagastuy-Brena, Aran Nayebi, Jacob Prince, Rosa Cao, and Daniel Yamins. Model-brain comparison using inter-animal transforms, 2025. URL `https://arxiv.org/abs/2510.02523`.

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`.

S1 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.

S2 Supplementary Discussion

Relation to prior benchmarks Our results replicate and 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. 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.

S3 Detailed Methods

S3.1 Alignment pattern similarity distributions

S3.1.1 Alignment pattern similarity distributions

To assess whether model-brain alignment pattern similarities fall within or outside the distribution of brain-brain alignment pattern similarities, we first define a subject-specific reference alignment pattern.

For a given ROI r and a subject t_0, we compute the mean brain-brain alignment pattern across all pairs of subjects (p, t) in which t_0 does not participate, i.e., $p \neq t_0$ and $t \neq t_0$. The subject-specific reference pattern for p_0 is then obtained by averaging over all such alignment patterns that exclude p_0:

$$\overline{\alpha}_r^{(t_0)} = \frac{1}{|P \setminus \{t_0\}| \cdot |T \setminus \{t_0\}|} \sum_{\substack{p \in P \setminus \{t_0\} \\ t \in T \setminus \{t_0\}}} \alpha_r(\phi_p, \Psi_t), \tag{S1}$$

where P and T denote the sets of all predictor and target subjects.

The brain-brain alignment pattern similarity distribution for subject t_0 , ROI r is then defined as the set of similarities between the reference pattern $\overline{\alpha}_r^{(t_0)}$ and all individual alignment patterns **in which t_0 functions as the target**:

$$\mathbf{D}_{\text{brain}}^{(t_0)} = \left\{ \rho\big(\overline{\alpha}_r^{(t_0)}, \alpha_r(\phi_p, \Psi_{t_0})\big) \;\; \forall p \in P \setminus \{t_0\} \right\}. \tag{S2}$$

Analogously, the model-brain alignment pattern similarity distribution for subject t_0 is computed using the model feature space ϕ_m as predictor:

$$\mathbf{D}_{\text{model}}^{(t_0)} = \left\{ \rho\big(\overline{\alpha}_{r_0}^{(t_0)}, \alpha_{r_0}(\phi_m, \Psi_{t_0})\big) \right\}. \tag{S3}$$

S3.2 Significance of fMRI-derived–to–structural alignment pattern similarity

To determine whether an APS value between a structural and an fMRI-derived alignment pattern is meaningful and not due to random chance, we create a null distribution of structural patterns, and compute APS between the fMRI-derived pattern to those random patterns.

We create random patterns by sampling 18 regions $\mathbf{k} = (k_1, \ldots, k_{16})$ at random from all regions contained in the full structural connectivity matrix $\tilde{C} = (\tilde{c}_{i,j})_{i,j=1\ldots M}$, $M > N$, containing additional regions to the ones included in our analysis. This yields one random alignment pattern per region,

$$\alpha_{rand,\mathbf{k}}(r) = [\tilde{c}_{\sigma(r),k_1}\ ,\ \ldots\ ,\ \tilde{c}_{\sigma(r),k_{16}}] \tag{S4}$$

where $\sigma(r)$ is the index of region r in matrix $\tilde{C}$. We then compute the APS to fMRI-derived alignment pattern $\alpha(\phi_p^r, \Psi_t)$ as

$$\rho(\alpha(\phi_p^r, \Psi_t), \alpha_{rand,\mathbf{k}}(r))$$

according to equation 3.5.4.

We repeat this 100 times to get a 95% percentile range of APS values due to random chance. We consider a fMRI-derived–to–structural APS value significant if it falls outside this range.

S4 Supplementary results

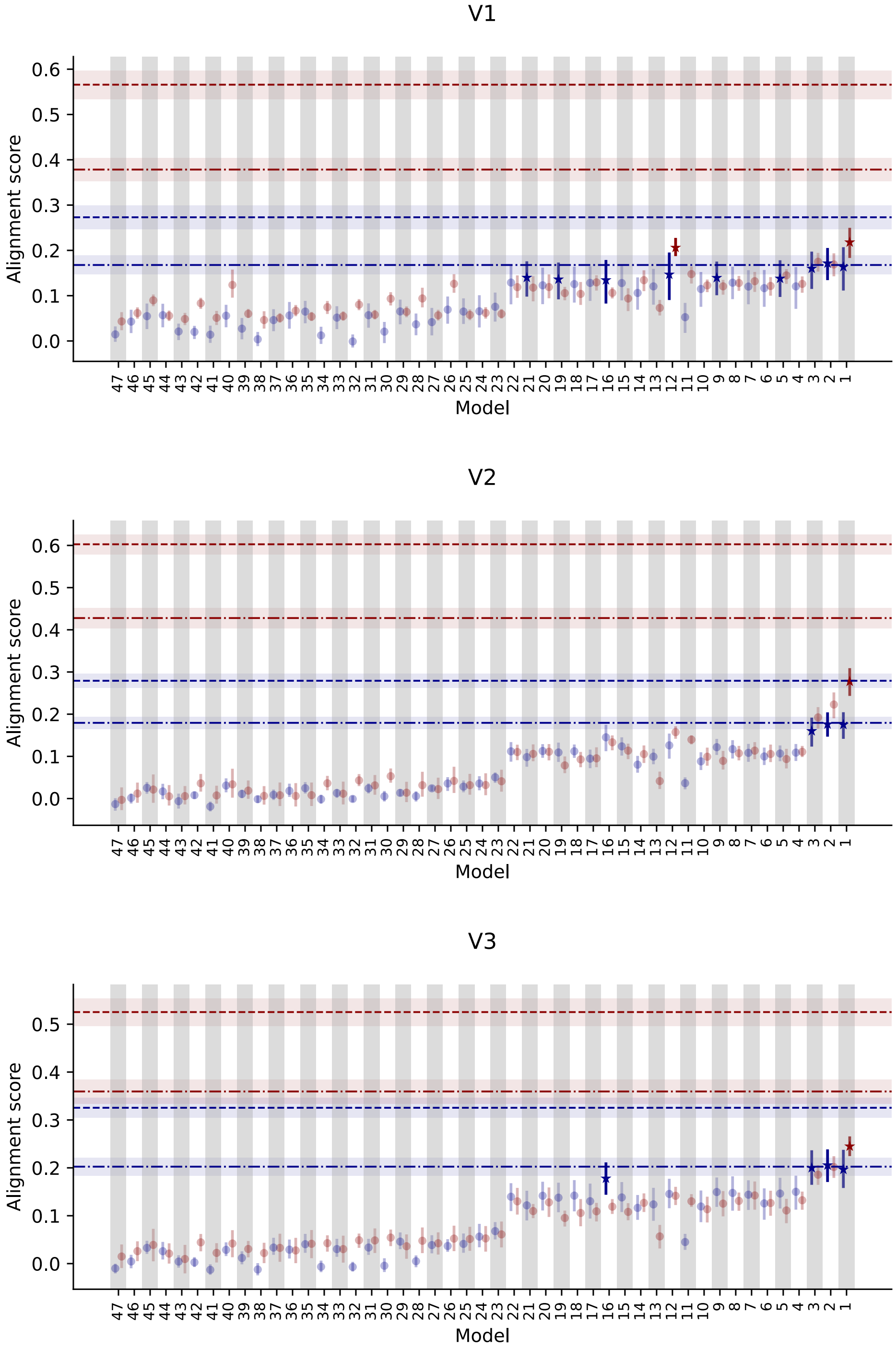

Figure S4.1: Benchmarking results for each ROI. Legend see main Figure 2

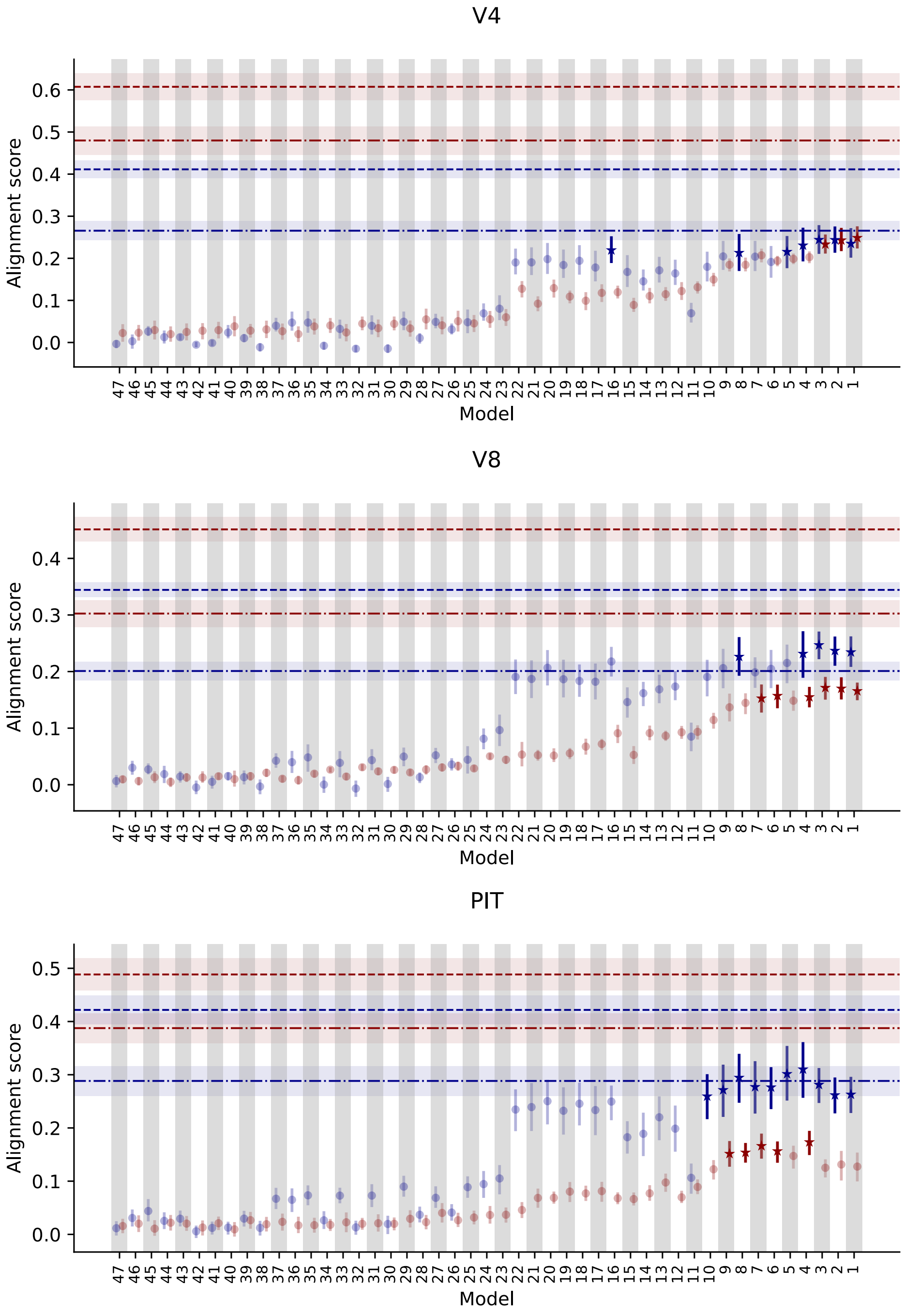

Figure S4.2: Benchmarking results for each ROI. Legend see main Figure 2

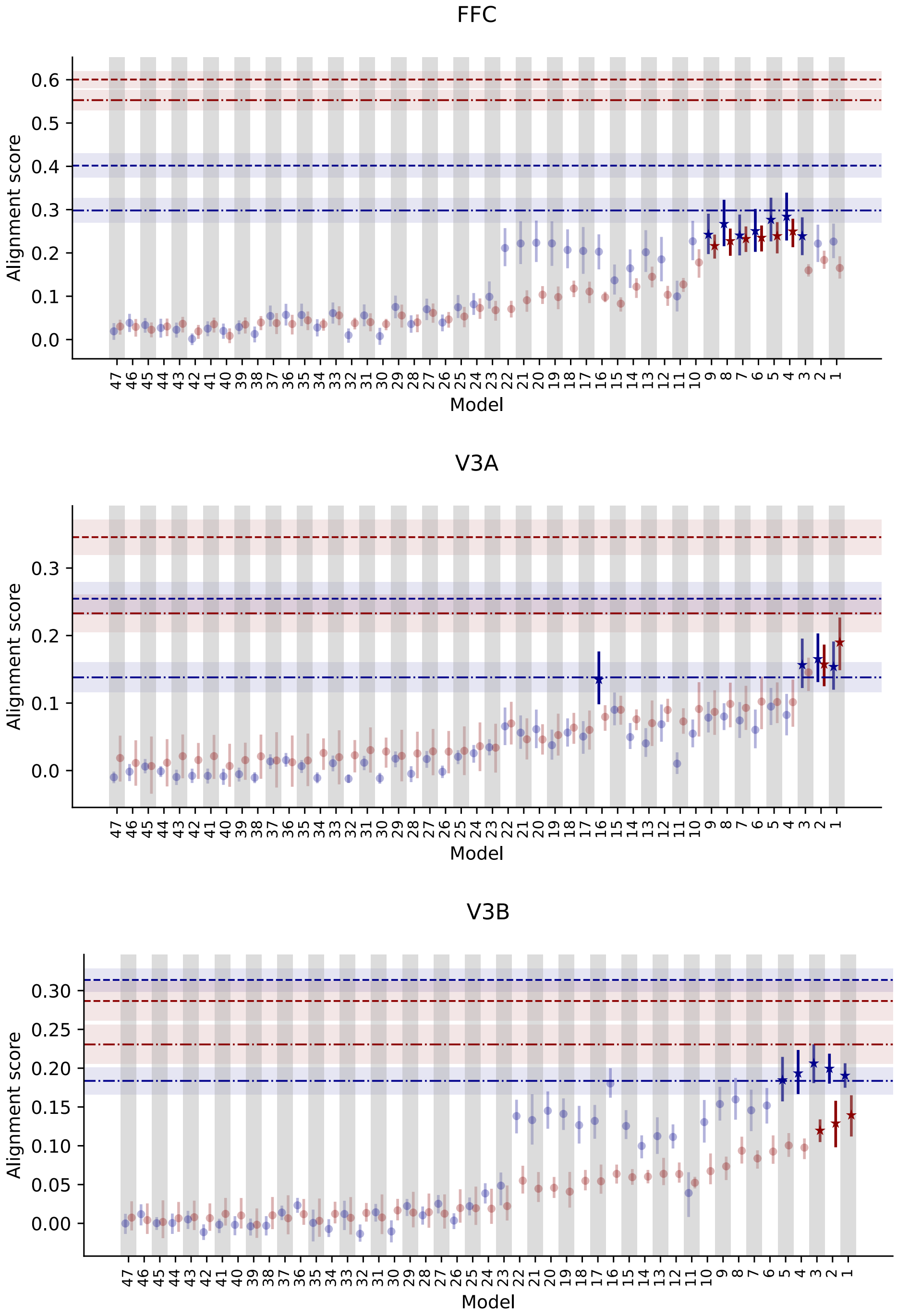

Figure S4.3: Benchmarking results for each ROI. Legend see main Figure 2

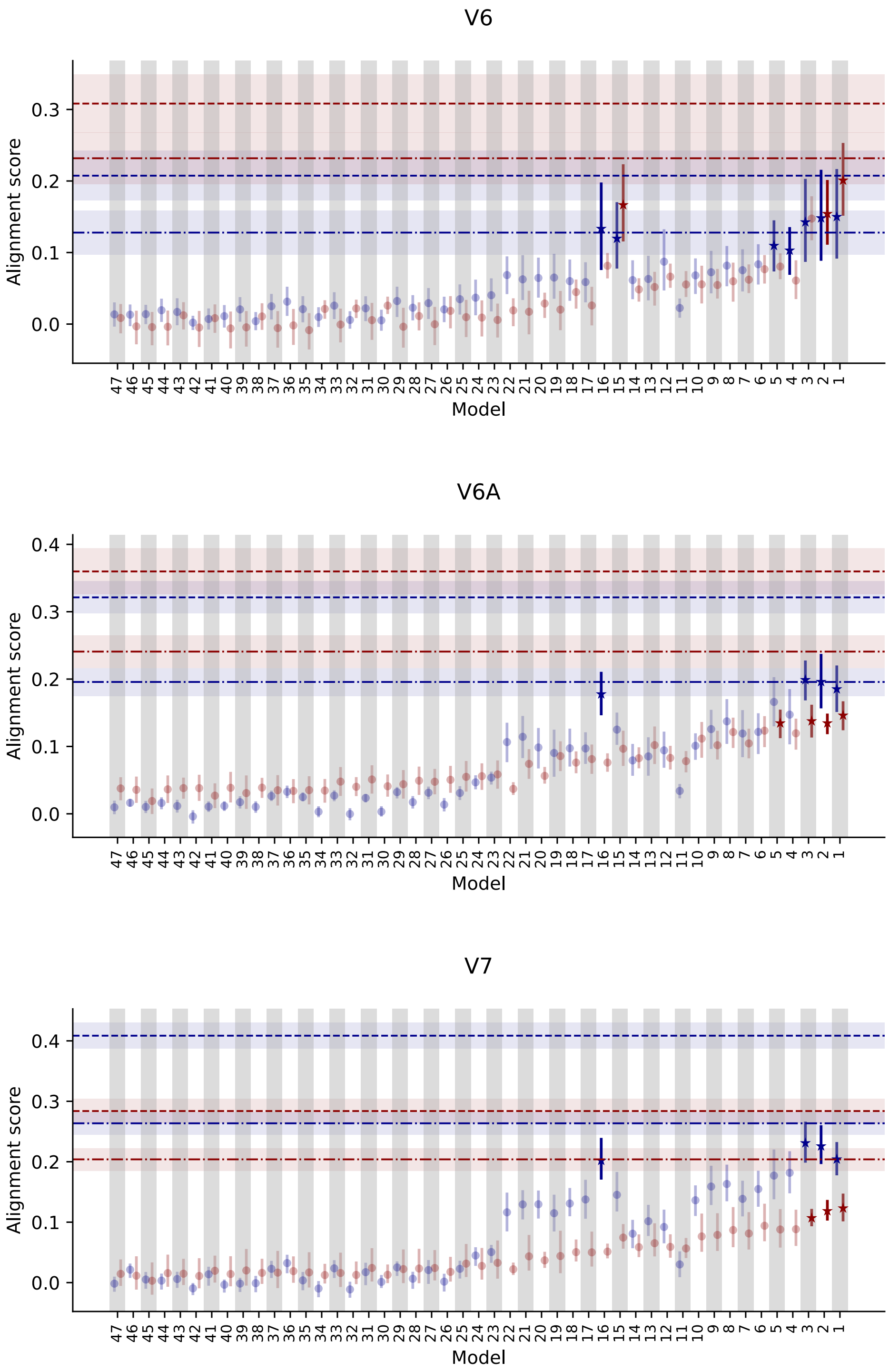

Figure S4.4: Benchmarking results for each ROI. Legend see main Figure 2

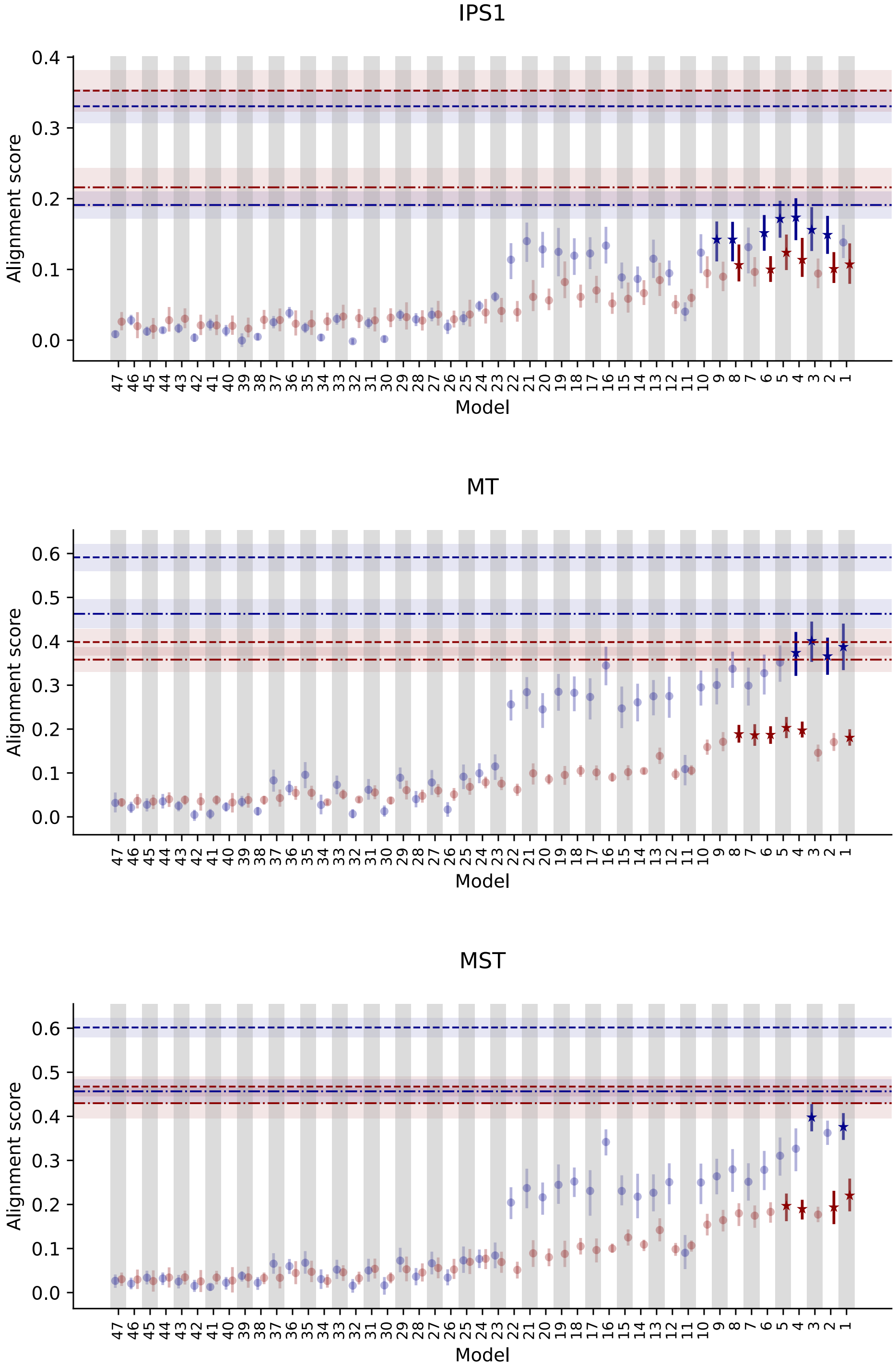

Figure S4.5: Benchmarking results for each ROI. Legend see main Figure 2

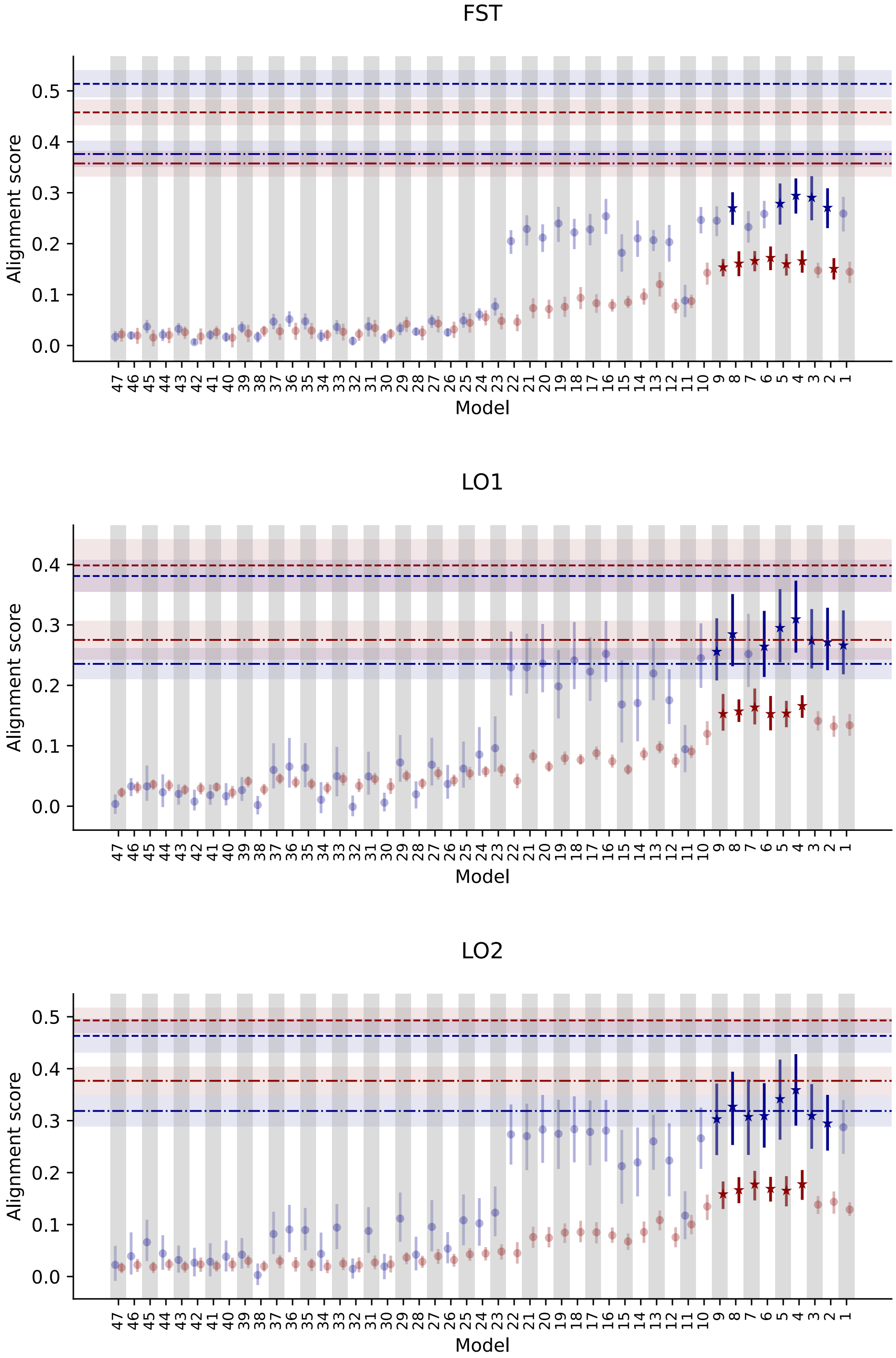

Figure S4.6: Benchmarking results for each ROI. Legend see main Figure 2

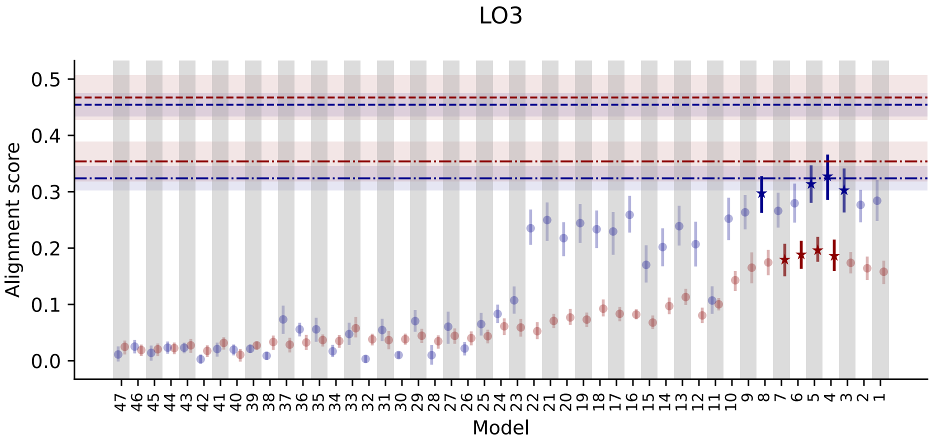

Figure S4.7: Benchmarking results for each ROI. Legend see main Figure 2

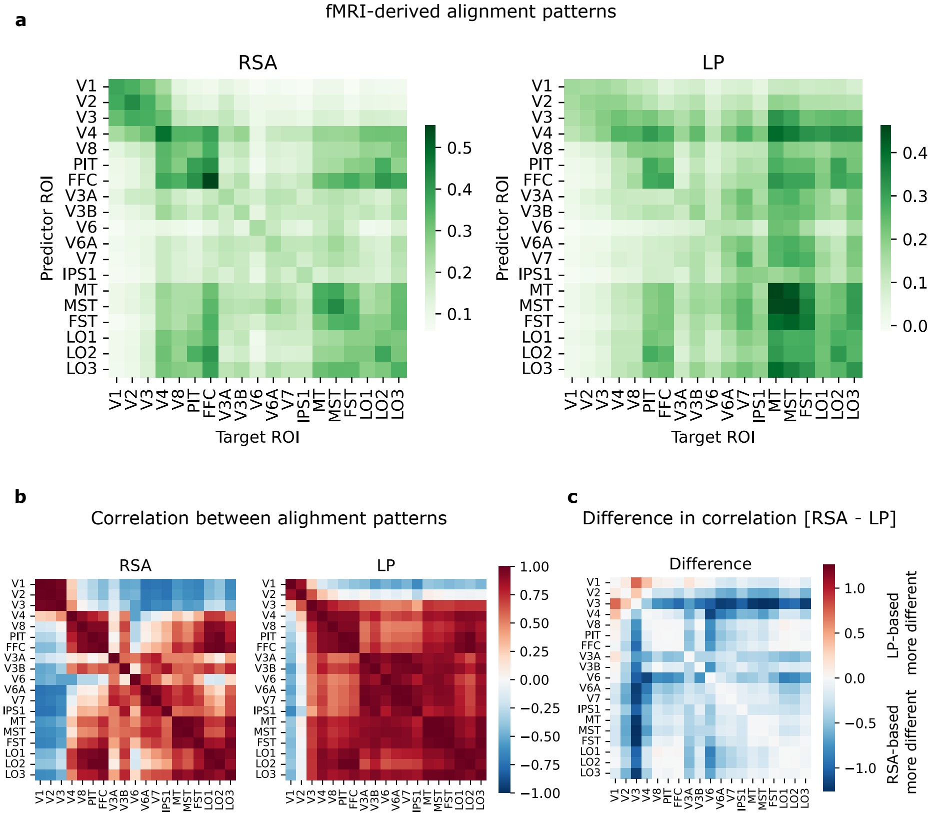

Figure S4.8: fMRI-derived alignment patterns and their similarities, RSA-based vs. LP-based. **(a)** Alignment patterns as heatmaps. **(b)** Confusions matrices based on pairwise correlations between alignment patterns. **(c)** Difference in correlation (RSA-titled panel in (b) - LP-titled panel in (b)).

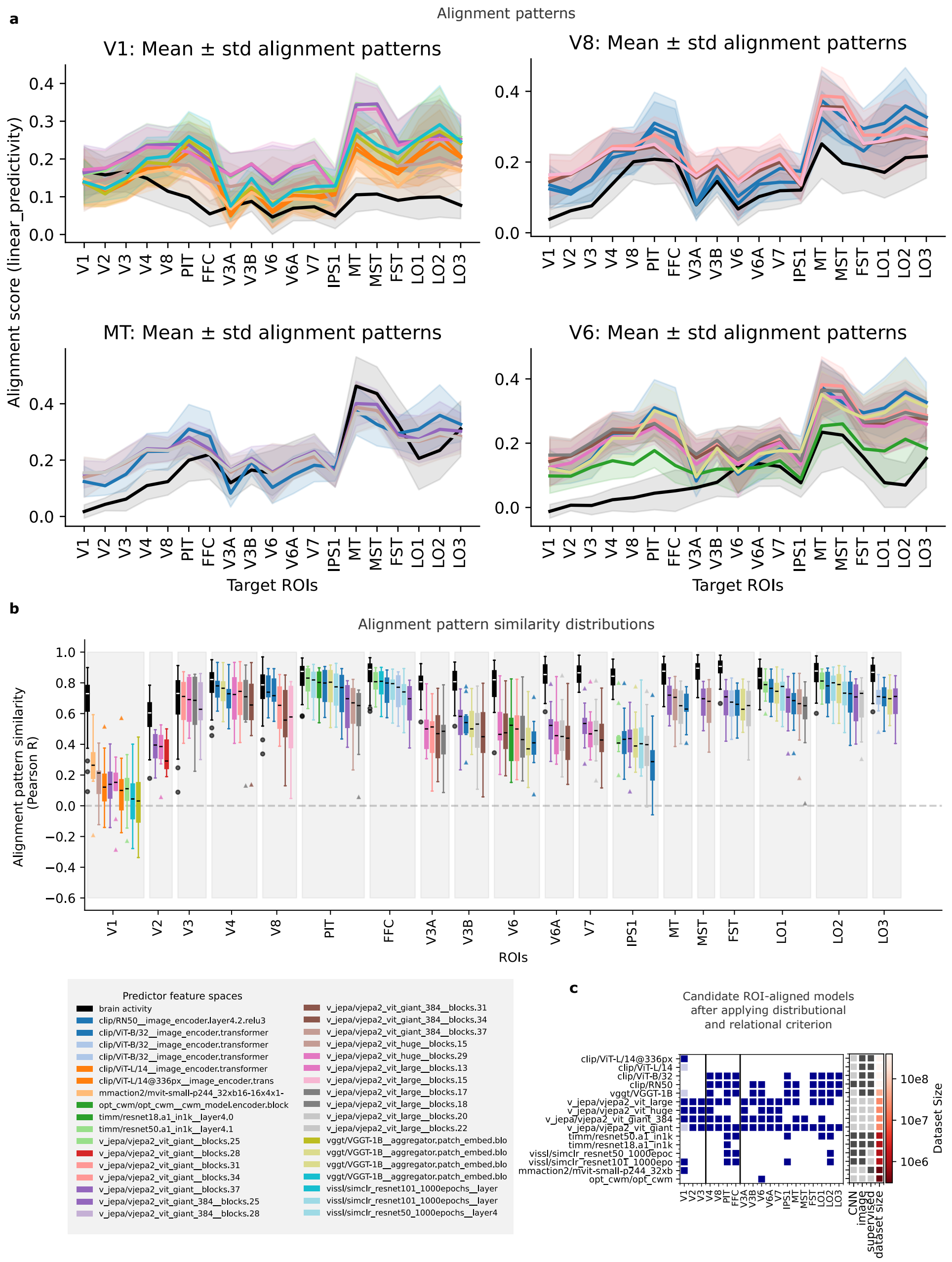

Figure S4.9: Alignment patterns and their similarities for ROIs and models, evaluated with LP. Same as figure 4 but with LP instead of RSA used as alignment measure.

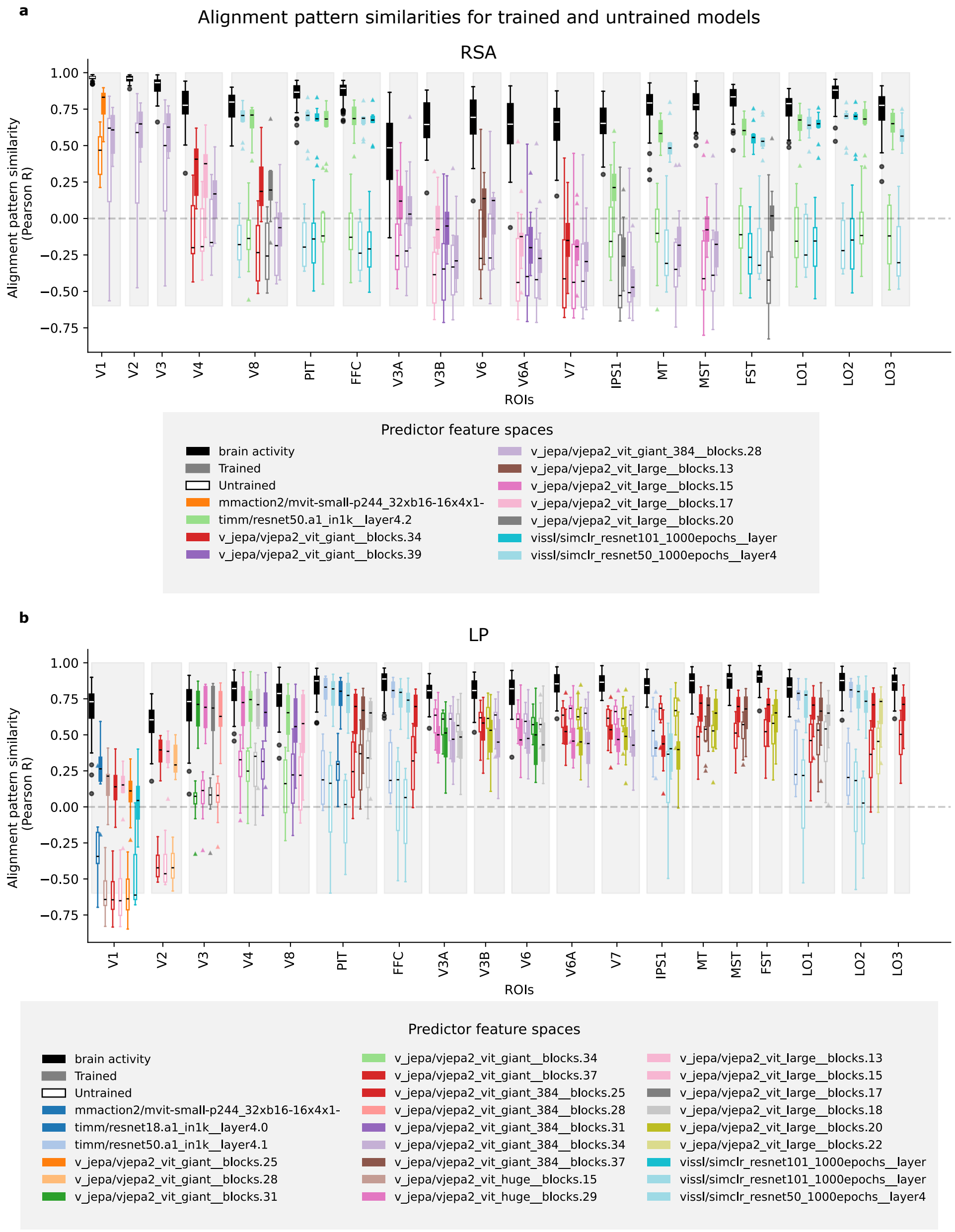

Figure S4.10: Alignment patterns and their similarities for trained and untrained models. a) RSA: Boxplots showing interquartile range of brain-brain APS (black) compared to model-brain APS for all models practically equivalent to the model with highest alignment to the region. Filled colorful boxes: trained models, white-filled boxes: untrained models. For model colors see legend. b) Same, but with LP as alignment measure.

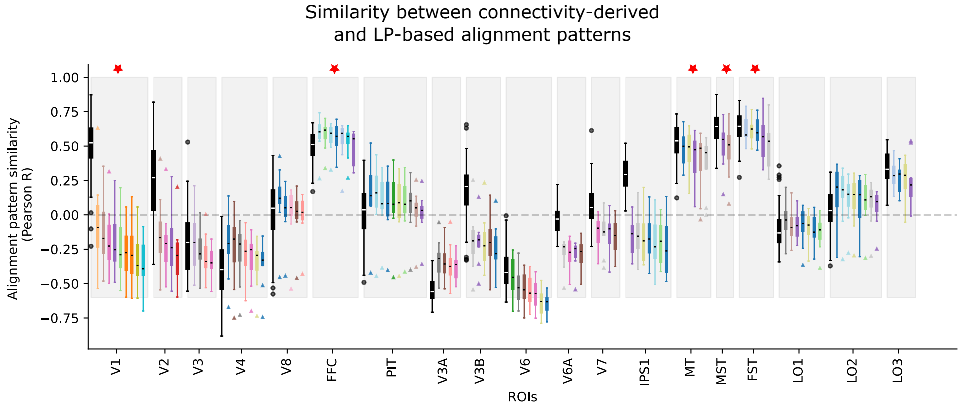

Figure S4.11: Similarity between connectivity-derived and LP-based alignment patterns. Same as fig. 5 but for LP instead of RSA.