TOWARDS INTERPRETABLE BRAIN-REGION SIMILARITY BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Comparing Deep Neural Network (DNN) representations to neural activity is a common way of evaluating computational models of the brain. Yet, design choices such as the alignment measure critically influence conclusions drawn from such benchmarks, which is a problem if we have no ground truth to validate a benchmark against. We propose to first sanity-check whether the benchmark is sensitive to the axes that differentiate the brain region under study, using *region matching* as a diagnostic test, and to require *feature specificity* of the similarity measure if the goal is to infer functional or mechanistic similarity from benchmark scores.

We evaluate ten similarity measures on both criteria using two large-scale fMRI datasets, the Natural Scenes Dataset and BOLD Moments. We find that region matching performance varies substantially by brain region and dataset. Region matching exposes preprocessing and measure implementation choices that lead to low benchmark sensitivity. For example, applying PCA impairs performance of some measures and using Pearson-correlation distance reduces matching accuracy with RSA. Further, the test surfaces that region identification on fMRI data strongly depends on the univariate signal component, a component that has questionable relevance for identifying functional similarity to DNNs. Also, many commonly used measures lack feature specificity: They cannot distinguish a region from overcomplete supersets or incomplete subsets of the region.

We recommend region matching as a sanity check before interpreting brain-DNN similarity scores, and to test for feature specificity when searching for precise model–region matches.

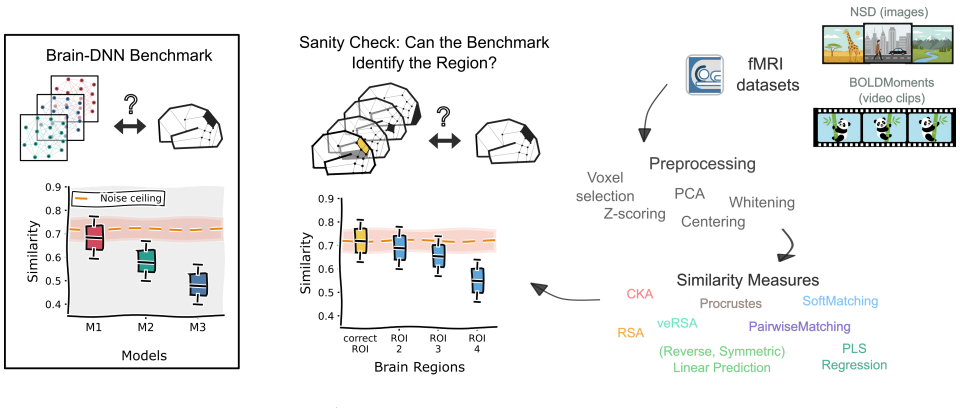


Figure 1: **Overview of our method.** We propose to gauge a similarity measure's discriminative capacity by using it to calculate the similarity between the neural activity in a target brain region in human A to the activity in all brain regions in human B. We then check whether the similarity between corresponding brain regions is indeed higher than the similarity between non-corresponding brain regions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

INTRODUCTION

Deep Neural Networks (DNNs) are increasingly used as models of the brain (Doerig et al., 2023; Yamins et al., 2014). One common approach is to evaluate representational similarity between DNNs and brain regions (Conwell et al., 2023; Finzi et al., 2023; Conwell et al., 2024; Ryskina et al., 2025; Sartzetaki et al., 2024; Tang et al., 2025), to find models that perform the same computation as the targeted brain region, possibly even by implementing a similar mechanism (Marr, 1982). The ultimate goal is to understand the brain by reverse engineering (Richards et al., 2019; Doerig et al., 2023). This is different from applications where neural predictivity is the goal itself, e.g., neural control, decoding or encoding models for use as engineering tools. We target pure neural alignment, of which predictivity is merely one candidate proxy.

Ideally, increasing representational alignment would point towards better models, on some objective one-dimensional scale. However, different similarity measures imply different model orderings in brain-similarity benchmarks (Soni et al., 2024; Huh et al., 2024), and many measures exist (Shah & Khosla, 2025; Khosla & Williams, 2024; Lin & Kriegeskorte, 2024; Kamiya et al., 2024; Ostrow et al., 2023; Kaniuth & Hebart, 2022; Williams et al., 2021; Nonaka et al., 2021; Ding et al., 2021; Kornblith et al., 2019)—see Klabunde et al. (2025) for a review of methods. Preprocessing and measure variants expand the benchmark design space further. Hence, a benchmark result depends on choices whose effect is rarely checked.

We suggest two criteria, a necessary one for benchmark settings, and one testing a measure property empirically. Both study the effectiveness of measures on the data modality that we want to model: brain data. The basic idea behind both is the following: If we had potential candidate models of a brain region, only one of which is a good model, which metric would consistently *recognize* the correct model? A major problem for validating brain-DNN similarity measures is the lack of ground truth – we do not *know* the best model of any brain region. But implicit in the search for models of a brain region is that we care about what this region has in common across subjects. Therefore, region A in one subject should be a good model of the same region in another subject, while a functionally distinct region B should be a worse model. Therefore, *region matching* is a natural sanity check for brain-DNN alignment benchmarks: Does the method consistently assign the highest similarity to the *same* region across subjects, compared to other regions?

While region matching has been used in the past (Lin & Kriegeskorte, 2024; Lahner et al., 2024; Thobani et al., 2025; Shah & Khosla, 2025), we extend existing work by conducting a principled and large-scale study across measures, datasets, and preprocessing methods and by providing recommendations for how to use and interpret it (section 3.1).

We find that seemingly innocuous preprocessing choices can have large detrimental effects on the sensitivity of individual similarity measures. We can further use region matching as an instrument to study what type of information a measure relies on to identify a region: Almost all measures lose accuracy when the univariate signal is removed, calling into question how pattern-based brain-DNN comparisons on fMRI data truly are.

We further propose to test a measure's *feature specificity*, to know whether it penalizes over- or under-completeness. A model matching the whole visual brain is not a correct model for region V1, nor does it tell us precisely what makes a model V1-like. Therefore, if the goal is to find region-specific models, or to find out what factors lead DNNs to be similar to a specific region, feature specificity is a desirable measure property.

To summarize, our main contributions are:

- We propose to use **region-matching as sanity check for brain-DNN benchmarks**, and to not interpret results on regions with low identifiability.
- We demonstrate its value empirically, showing strong variations of region identification, in a large-scale evaluation across measures and preprocessing methods on two fMRI datasets, and brain areas spanning posterior visual cortex and association regions.
- We surface that robust region identification on these two datasets relies on the univariate signal, across the ten measures in our study.
- We introduce **feature specificity** as a property of a measure, as well as a way to determine its absence, narrowing down the choice of measures depending on the research question.

- We show that commonly used measures are not feature specific, so that they are unlikely to rate a precisely matching model as more similar than an all-encompassing, unspecific one.

We hope that our findings help researchers narrow down which measure of similarity to use in which context, and that region-matching will become a standard check for brain-alignment benchmarks.

Code will be made open-source upon publication.

Related Work

Benchmarking similarity measures. There have been numerous attempts to validate neural similarity measures: They have been correlated with functional similarity (Bo et al., 2024; Ding et al., 2021), evaluated for their capacity to differentiate random from trained networks (Bo et al., 2024; Ding et al., 2021; AlKhamissi et al., 2025) and scrambled from well-formed inputs (AlKhamissi et al., 2025). Further, they were tested on their ability to differentiate DNN architectures or layers (Bo et al., 2024; Han et al., 2023; Ding et al., 2021), robustness to DNN random seed or stimulus selection (Han et al., 2023; Ding et al., 2021; Lin & Kriegeskorte, 2024), whether they peak at the most brain-like point of topographic structure of a DNN trained with a spatial constraint regularizer (Soni et al., 2024), and whether optimizing artificial *feature maps* to maximize a measure will transfer known ground-truth variables to the features (Cloos et al., 2024).
While each validation check contributes a puzzle piece of what we might expect of a 'good' measure, many can be argued with—for example, functional correspondence requires that we already have a very good idea of the task that a brain area performs. We are left with a pool of evidence to weight subjectively, and the question of how to choose a measure remains without clear answer.

Similarity measures do not agree. Different functional and representational similarity measures lead to different orderings of DNNs as models of different brain regions (Soni et al., 2024; Ahlert et al., 2024) and to different conclusions (e.g., (Huh et al., 2024) Fig. 13). Measure correlations range from large to small negative values (Soni et al., 2024; Bo et al., 2024).

Region identification accuracy as validation tool. The test that we call region matching has been used in the past, to evaluate different types of Representational Similarity Analysis (RSA) (Lin & Kriegeskorte, 2024), to validate datasets (Lahner et al., 2024; Finzi et al., 2023) on distinguishability of visual streams, by Shah & Khosla (2025) to validate their measure (see also section 5), and by Thobani et al. (2025) using the silhouette score and comparing versions of Soft Matching, RSA and linear predictivity (LP). Our results based on different statistical methods partially contradict theirs (see also section K.3). Thobani et al. (2025) use the term *specificity* to refer to what is measured by region matching, which is distinct from *feature specificity*.
Finzi et al. (2023) compare RSA and linear predictivity scores within and between three cortical streams, to validate dataset quality of the NSD (Fig. S2 b). Our finer-grained region selection allows to detect some regions that are not identifiable. (Kornblith et al., 2019) propose a sanity check based on a similar idea, testing CKA, LP and three CCA variants on a DNN-layer matching task, while we test on the brain side, which is the side we are trying to explain in brain-DNN alignment. We refine existing approaches through addressing the problem of subject-specific anatomic region dislocation, presenting guidelines how to interpret it, proposing to analyze the resulting data statistically by use of Cohen's d_z, and by evaluating ten different types of similarity measures, and multiple combinations of preprocessing steps.

Critical perspectives on brain-DNN similarity benchmarks. Brain-DNN similarity measurement has been called into question both conceptually as well as empirically: (Kapoor et al., 2025) show that a networks' neural Procrustes-alignment is often already maximized after a single epoch of training. (Guest & Martin, 2023) warn that correlational brain-DNN alignment is insufficient evidence for mechanistic similarity, as different mechanisms can produce similar outputs. (Bowers et al., 2023) draw attention to the discrepancy between claims of alignment between brains and DNNs, to the failure of DNNs to reproduce a range of human perceptual phenomena in controlled psychological experiments. Others have noted the potential of representational alignment (Sucholutsky et al., 2023), calling for more knowledge transfer between disciplines and proposing a joint terminology. They also stress the importance of exactly specifying preprocessing, metric implementation and the data used, and list measure and dataset choice as open problems.

METHODS

A REGION MATCHING SANITY CHECK

We propose to use region matching to test whether a benchmark is sensitive to a target region. Region matching asks: Does the similarity measure S used by the benchmark assign the highest similarity to the matching region across subjects, compared to the similarity to other regions that we consider meaningfully different? This sanity check reuses the benchmark setting and only requires the (train) data used by the benchmark. It does not test the measure in isolation: It tests the joint benchmark setting, consisting of *dataset, preprocessing, similarity measure* and *target region*. We provide a summary of this section in box A.

Notation We assume the benchmark defines brain activity feature maps $X_r^{s_i}$ for regions $r = 1, ..., R$ measured for subjects $s_1, ..., s_M$. A similarity measure S assigns similarities to pairs of such feature maps: $S(X_r^{s_i}, X_{r'}^{s_j})$. Similarity measures can be asymmetric; we distinguish *source* and *target* inputs to S, where the first argument is the source, the second the target. For example, in a benchmark using linear predictivity as measure, $S(M, X_r^s)$ would be the prediction score of model M, predicting brain region r in subject s.

Interpreting region matching as performance on an ideal model zoo Region matching can be thought of as a test how a benchmark performs in an ideal setting, on a set of brain-like 'models', each closely matching one of the target regions. We can use another subject's data as matching 'model', or a suitable aggregate across subjects. We here average fMRI beta values across the remaining subjects. Single subject comparisons perform worse (figs. 48 and 49).

We propose two interpretations that inform how to quantify region matching. These interpretations also suggest what numeric values to consider as a success or a failure of the sanity check.

First, conservatively we can consider each target subject as a single benchmark for a region, leading to a binary accuracy per subject:

$$\mathrm{acc}_{r,s}(S) = \begin{cases} 1, & \text{if } S(\tilde{X}_r, X_r^s) > S(\tilde{X}_{r'}, X_r^s) \quad \forall\, r' \in \text{distractors}(r), \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Here, $\tilde{X}_r$ can be another subjects' feature map $X_r^{s_j}$, $j \neq i$, or a feature map derived from multiple subjects' data of region r.

Averaging across subjects, region r's **accuracy**, $\mathrm{acc}_r(S) = mean_s(\mathrm{acc}_{r,s})$ can thus be interpreted as the estimated probability that a single benchmark instance determines the correct model of the region. That is, an accuracy of 60% would imply a failure probability of 40%, in this conservative interpretation. Accuracy as quantification has the added advantage that it is invariant under any monotonous transform of S: Whether a measure returns Pearson r or an R^2 score, they can be compared on equal footing.

However, benchmarks are usually evaluated on multiple subjects. Given how a benchmark picks the best-matching model and tests differences between models, we could instead compute statistics based on the full set of subjects, to estimate the success of a hypothetical new instantiation of the multi-subject benchmark, on a similar ideal model zoo. This, however, requires benchmark-specific statistical treatment. In the following, we use **Cohen's $\mathbf{d_z}$** (Cohen, 1992) to the closest distractor (as described in section G.1.2) as one such population-based statistic. It corresponds to a benchmark that uses the mean similarity across subjects to rank models (Sartzetaki et al., 2024; Schrimpf et al., 2018; Conwell et al., 2024), and that computes significance of differences with a paired t-test. For example, with 10 subjects, the 75%-prediction interval for the mean difference on a new set of subjects would exclude zero if $d_z \gtrsim 0.55$. Cohen's d_z is invariant to affine but not to nonlinear transforms of S. Accuracy and Cohen's d_z broadly share decision boundaries (figs. 46 and 47). We note that in both accuracy and Cohen's d_z, comparisons are between similarities of multiple sources to the same target, obviating the need for noise ceiling normalization.

Distractors and region granularity The choice of distractor regions determines the resolution we expect of the benchmark. For example, our experiments exclude direct neighbor regions, to account for anatomical shifts of a region between subjects - we do not require our anatomically defined

benchmarks to tell apart adjacent regions. More about our specific distractor region choice can be found in G.3. If region matching performance is low across regions, and choosing coarser target regions improves it, running a modified benchmark with these coarser regions better reflects what the benchmark can differentiate. Also, if a region is frequently confused with a specific other region, merging them might yield stronger region matching performance and thus a better identifiable target region for the benchmark.

Limitations Region matching has blind spots, limiting it to the role of a sanity check. First, region matching is susceptible to shortcuts. A simple example is feature map size: A measure matching by voxel count could achieve high accuracy in a naive test. We control for this in section I.2 and section I.1. More generally, the measures can pick up on confounders that are not related to neural activity Hebart & Baker (2018). Second, biases that take effect only in DNNs will not be visible in region matching performance. An example is CKA, which rewards larger dimensionality (Murphy et al., 2024). When DNNs' dimensionality varies along a much larger range than brain data, this bias might drive brain-DNN similarity scores, while being invisible in brain-to-brain comparisons.

Region matching is necessary and complementary to other criteria Region matching tests how sensitive a measure is to the empirically observed feature differences between regions. If such sensitivity is lacking, the benchmark has nothing to compare to in DNNs. Sufficient region matching performance is therefore a necessary, but not sufficient criterion. It has to be satisfied in addition, and not in trade-off, with other criteria.

FEATURE SPECIFICITY

Let us assume that we want to narrow down which aspects of the architecture, training data, loss, or position inside the network, increase similarity of a DNN layer to a brain region. Then it might be the case that an architecture is flexible enough to support submodules that fit the region, or that one loss function benefits from a second, intermediate loss which is in fact the true loss optimized by the region. Here, we want to distinguish an model that contains a fitting subcomponent, from one that fits the region exactly. Otherwise our search might stop early at the over-complete system, missing the specific aspect that makes a layer region-like.

We propose to call a measure *feature specific* if it is sensitive to addition and removal of features. That is: does its similarity score drop when unrelated features are appended, or when some of the matching feature dimensions are removed? Feature specificity is also a prerequisite for claims of mechanistic or functional similarity, since a model matching the region's features plus many others, which implement different functionality, is not performing that region's computation. A summary of feature specificity and how to test for it is given in box A.

We note that lack of feature specificity is expected for all measures where adding a feature dimension is expected to only increase, not decrease scores. This includes linear predictivity, pairwise matching, PLSReg and also veRSA.

Testing feature specificity We test feature specificity using a variant of region matching, by asking: Can the benchmark differentiate a region composed of multiple regions, from its constituents? We evaluate whether appending features from another region R reduces similarity for a correct match for region r:

$$S(X_r^{s_j}, X_r^{s_i}) \overset{?}{>} S([X_r^{s_j}, X_R^{s_j}], X_r^{s_i})$$

And for testing detection of feature removal, we look at a joint region composed of r and R: Does dropping R reduce the similarity compared to the correct match?

$$S([X_r^{s_j}, X_R^{s_j}], [X_r^{s_i}, X_R^{s_i}]) \overset{?}{>} S(X_r^{s_j}, [X_r^{s_i}, X_R^{s_i}])$$

To be certain that region R contains features that are functionally different from those of r, we use an extreme comparison: R is composed of the 14 (BMD) or 13 (NSD) remaining regions except r. Detailed considerations what the added and dropped features must satisfy are given in section B,

We report how feature specific a measure appears, as the *self-preference rate*, which is the fraction of subjects s_i and regions r where both inequalities are fulfilled simultaneously - excluding regions r

that are clearly unidentifiable ($min(d_z) < 0$ in region matching). We use a lenient threshold here because of the large regions we compare to. We note that testing feature specificity empirically allows to detect a *lack* of feature specificity, not to show that a measure is feature specific.

Limitation: Feature specific measures applied to partially recorded regions When the benchmark data sub-samples a region - often the case in electrophysiology data - a feature specific measure will also penalize features that would contribute to a match with the full regions' features, but do not match with the actually recorded subset. It is not known how this penalty trades off with quality of the match to the subset.

Experimental details

Datasets BMD (10 subjects) and NSD (8 subjects) are described in section D, similarity measures in section C, and our checkerboard region choice of 15 (BMD) and 14 (NSD) regions in section G.3. We test the effect of various factors on region matching, including whitening (details in section F.1), principal component analysis (PCA)-based feature reduction to the first 50 components, z-scoring, and for RSA additionally centering and different distance functions, which we also ablate for veRSA. Our default setting is to z-score only, implying centering. We also drop voxels below a noise threshold of 0.2 NCSNR (section E), which we ablate in section F.2. Univariate signal ablation is described in section 4.1.

Results

Region matching performance depends on measure, region and dataset

We show fig. 2 that some measures seem overall less discerning, but none dominates consistently across regions. Popular measures like linear predictivity (LP) and Pearson-correlation based RSA (P-RSA) identify fewer regions than SoftMatching or Euclidean RSA (figs. 13 and 14). Aggregating across our region set fig. 3, we find that standard measures like P-RSA and LP ranking mid-field, with average identification accuracy of 0.6 or below. Reverse and symmetric LP show worst identification. While reverse linear predictivity is not typically used as standalone measure, symmetric linear predictivity symmetrizes LP and improves its feature specificity (see section 4.2) – which, as we show here, comes at a cost. We also test whether measures disagree less when comparing them on regions they identify section H.3, and find that they disagree even then, even when comparing brain data without modality gap.

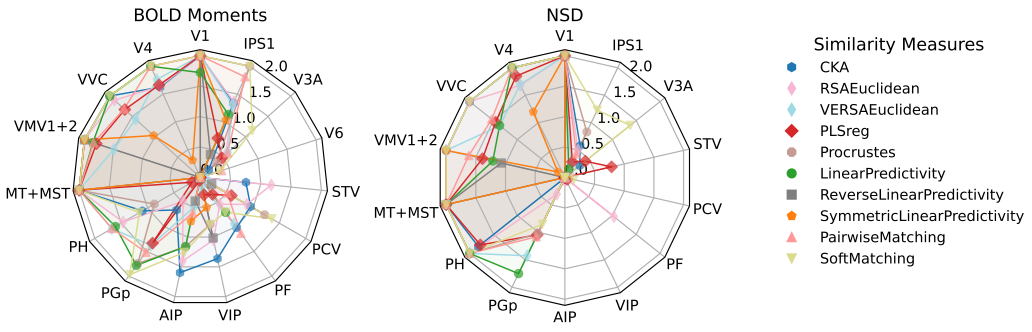


Figure 2: **Discernibility by brain region.** Shown is minimum Cohen's d_z across distractor ROIs, for each ROI and similarity measure. **Left: BMD.** Most measures perform well on early visual and ventral (VVC, VMV1+2) regions as well as on dorsal regions MT+MST, IPS1 and PH. **Right: NSD.** The difficulty of identifying higher-level association areas, with the exception of PGp, is more pronounced. Note V6 is excluded on NSD.

Identifiability of a region cannot generally be inferred from the noise level in a region (section E). This shows that sensitivity to what identifies a region is not guaranteed by standard measure choices or noise-based region rejection, so that a test like region matching is necessary.

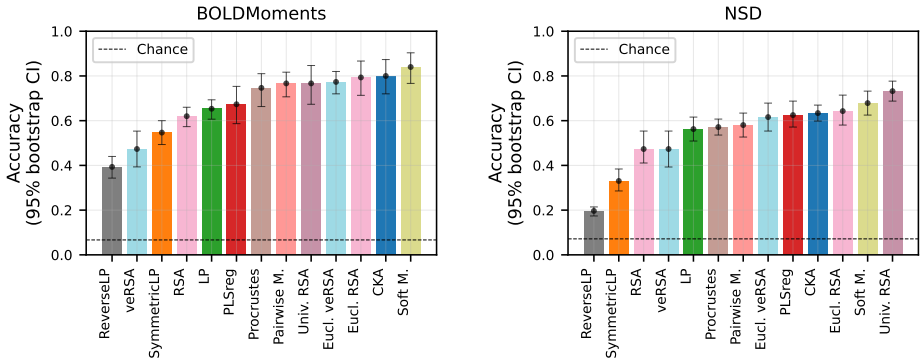


Figure 3: **ROI matching accuracy.** We plot each measure's ROI matching accuracy, averaged across subjects, i.e., $mean_r(\text{acc}_r(S))$, on a checkerboard-patterned subset of regions, for BMD (left) and NSD (right), across all measures tested. Standard setting. Error bars show 95% bootstrap CIs across target subjects. Per-measure CIs are shown in fig. 46.

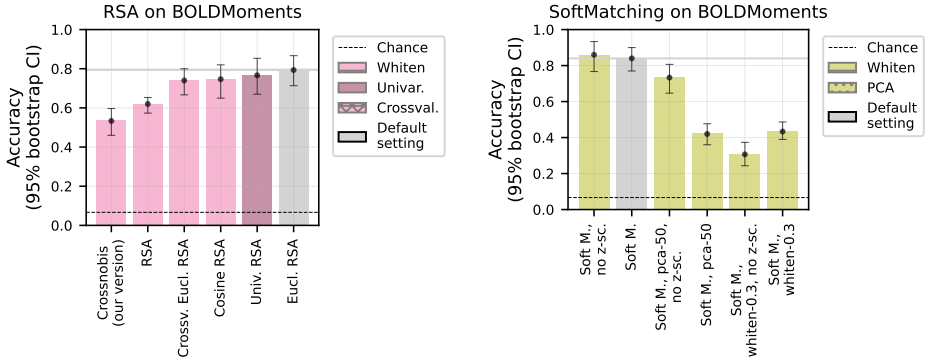


Figure 4: **Common preprocessing choices have detrimental effects on RSA and Soft-matching region matching scores.** Region matching on BMD, RSA variants (left) and SoftMatching with different preprocessing (right). For SoftMatching we apply re-normalization after PCA or whitening transform. Default setting (gray) always means: z-scoring, no feature reduction beyond removal of noisy voxels (section F.2).

Preprocessing and measure details affect identifiability

We show the impact of different preprocessing choices on aggregate region matching accuracy in fig. 4, and in figs. 16 to 18 for all measures. Common design choices like PCA-based feature reduction strongly reduce discriminative capability for some measures, but not for others. For example SoftMatching is strongly impaired by PCA and also by multivariate noise normalization (whitening).

Given its role in fMRI analysis, we test multiple variants or RSA, including whitened, crossvalidated Euclidean RSA (Crossnobis, (Walther et al., 2016)), and a 'univariate' variant that compares scalar region-average activation between stimuli (section C). Dropping de-meaning (Ramírez, 2017) has no positive effect here (fig. 18). Crossnobis shows comparatively low region identification, worse than using either crossvalidation or whitening alone (fig. 20). P-RSA also shows decreased identification, in contrast to univariate RSA, a measure that ignores the pattern of activity across voxels. A similar pattern is visible on NSD and for veRSA variants (fig. 19). This reduced sensitivity reflects a more general trend, wherein the univariate signal component seems required and sufficient for a reasonably high region matching accuracy, as we show next.

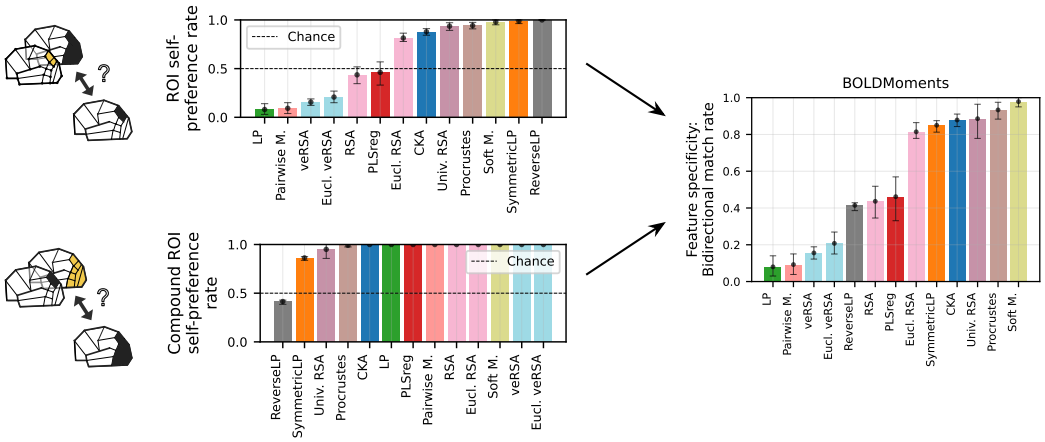


Figure 5: **Feature Specificity: Can the measure detect a precise match?** See also section 3.2. **Top left:** Fraction of subjects and regions where same-ROI similarity exceeded similarity to the compound ROI. This estimates for how well a measure can detect auxiliary features. **Bottom left:** Fraction of subjects and (subset) regions for which the similarity of the compound ROI to itself exceeded that of the compound ROI to the small ROI, as estimate for how well a measure can detect missing features. **Right:** Bidirectional match rate, the fraction of (subject, ROI) pairs where both directional matches are fulfilled jointly. Corresponds to a logical elementwise *and* of small-ROI and compound-ROI self-preference rate, before aggregation. All plots are for BMD. Results for NSD are shown in section H.2 . Errorbars show 95% confidence intervals from bootstrapping over subjects.

Beyond sanity checking: Region matching shows univariate signal reliance

The univariate signal is, informally, how much a region responds to a given stimulus. We compute it as a regions' mean response per stimulus (i.e., averaged across voxels). Pearson-correlating two stimuli's voxel patterns drops this per-stimulus mean, while univariate RSA uses it as sole input. Univariate information is classically used to identify regions in fMRI data (Kanwisher et al., 1997; Epstein & Kanwisher, 1998; Saxe et al., 2006), but the insights it provides about functioning of a region are limited, beyond informing about the kinds of stimuli that the region appears to process.

We evaluate region matching with the univariate component subtracted from the data - so that for every stimulus and every region, the average activity is zero. This tests region matching based on pattern information alone (fig. 28, center). Accuracy decreases across measures, with the exception of P-RSA. We show in fig. 27 how this affects identifiability per region. For BMD, with the univariate signal dropped, 11 of 15 regions stay above a $min(d_z)$ of 0.5 for at least one measure, compared to 14 of 15 in the standard setting. That is, four regions appear only identifiable with their univariate response present. In contrast, when using only the scalar univariate signal, region matching is close to or even more accurate (in particular for NSD) than with the full signal (fig. 28, left).

Feature Specificity: Can a measure detect superfluous or missing features?

Feature specificity results are shown in fig. 5 (section H.2 for NSD), jointly (right) and disassembled into detection rate of superfluous (top) and missing (bottom) feature dimensions. As expected, measures that reward over- or under-completeness show low feature specificity. More surprisingly, P-RSA and symmetric LP also do not reliably distinguish a large compound region spanning visually driven cortex, from each of its 15 constituents (14 for NSD). Meanwhile, Soft Matching and Procrustes appear feature specific on these two datasets. However, with the univariate signal dropped, no measure remains feature specific (section H.1.6). We note that given the extreme size difference between the regions that are being compared, even a self-preference rate of 80% indicates a partial lack of feature specificity.

DISCUSSION

Conclusion How to compute similarity between brain data and DNN features is an unsolved problem. We take a step back from this fundamental question and show that brain-DNN benchmarks need a sanity check, to ensure they are sufficiently sensitive to the regions for which they aim to find models. One misguided preprocessing choice can severely impair benchmark sensitivity. We provide guidelines for how to implement a region matching sanity check, and how to statistically evaluate and act on results.

An orthogonal concern are (in)sensitivities to DNN-specific axes of variation, such as architecture, training stage or layer depth. On their own, these are neither good nor bad for brain-DNN similarity comparisons, because the relevance of each such model difference to brain similarity is not known.

We here take two baby-steps towards the fundamental question what measure and benchmark to choose: First, we demonstrate how region matching can be used as an instrument to find out what features a benchmark is sensitive to: Ablating the univariate and multivariate signal components, we show that identification on the two fMRI datasets we study relies on the univariate component for many regions. Further, it is questionable that the complex inner workings of a brain region can be identified by its mean activity alone, hinting at a resolution limitation of fMRI-based brain-DNN alignment for gaining insights about the brain. However, since region matching always tests a joint benchmark setting, we cannot exclude that non-anatomical region definitions or a different measure or preprocessing choice can preserve identifiability under univariate removal.

Second, we introduce *feature specificity* to capture the sensitivity of a measure to additional and missing features. A benchmark based on a less feature specific measure will not differentiate between an all-encompassing model, and one capturing the exact processing of a region. Inferences drawn from model-region matches can only be as region specific as the match itself, be it claims of mechanistic similarity, or about what loss or architecture make a model similar to a region. Therefore, knowing that a measure lacks feature specificity can inform which research questions we can expect to address with it.

Limitations Region-matching is a region- and dataset-specific sanity check. It cannot be used as meta-benchmark to determine the 'best' measure to use, since it does not evaluate the measure in isolation. Nor can it be used to determine the 'best' benchmark, because of blind spots and possibility of shortcuts (section 3.1), and because it fundamentally does not tell us what type of similarity, if any, best reflects functional or mechanistic correspondence between the systems of which the features are intermediates. Region matching ablations also do not inform about suboptimal *DNN* feature preprocessing, as information is likely encoded differently in DNNs compared to fMRI data.

Similarly, conclusions about the feature specificity of an alignment measure drawn from two fMRI datasets might not transfer to other modalities. We use two datasets to estimate how results generalize across (fMRI) datasets. While this shows a mostly consistent set of highly feature selective measures, further experiments on a larger variety of brain data are needed to solidify our knowledge of the feature specificity of a measure. Feature specificity in a DNN-like feature space would be equally relevant, but considerably more involved, because DNNs' random initialization creates random features that might survive training even if they are not functionally used by the network. That is, it is harder to ensure that one set of DNN features does not contain a good match for another set of DNN features. We also note that our feature specificity test adds and removes a very large set of features, which might miss a subtler lack of feature specificity.

Furthermore, region matching rests on other subjects' data being a 'good model' of a region. Noise means that the derived 'model' is likely somewhat suboptimal, and region matching is therefore conservative. For electrophysiology, where recordings typically sample only part of a region, the problem can be severe, such as when different subjects' recordings do not overlap. For such cases, also, feature specificity cannot be a requirement because coverage of region features is by nature incomplete.

AI USE STATEMENT

In this work, we used generative AI tools to help develop conceptual models (in particular, as rubber duck for discussing feature specificity and region matching), to help formulate mathematical claims, to refine and refute hypotheses, to design and provide feedback on research methodology and experiments (such as to select and refine statistical methods), to interpret results, to adjust code to create uncentered versions of BMD and NSD, and to translate terms and concepts for writing. We have not used generative AI tools to propose hypotheses or for qualitative data analyses. Generation of synthetic data sets, mathematical proofs, theoretical models beyond conceptual frameworks, cleaning and reformatting datasets, and thematic data analyses are not applicable to this work.

We further used generative AI to help create schematics and images that are part of Fig. 1, to create and edit software code, to summarize and analyse existing literature, to identify or narrow down research gaps, for brainstorming and rubber-ducking generally, for searching dataset files such as the Glasser-atlas probabilistic maps, to identify relevant literature, to discuss and brainstorm the best choice of formulations, including naming of feature specificity and its constituents, and we have used it to simulate additional feedback. We have also used it to create first versions of, and to improve, LaTeX typesetting for equations and figures.

We have reviewed all AI-assisted work. We have made heavy use of LLM assisted coding, but humans have reviewed all parts of the code and humans were involved in writing most of the code. We have checked related work that AI recommended, and we have ensured mathematical derivations and, to our best understanding and knowledge, statistical methods, are correct. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We will make code open source upon publication to ensure reproducibility.

REFERENCES

Jannis Ahlert, Thomas Klein, Felix Wichmann, and Robert Geirhos. How aligned are different alignment metrics? *arXiv preprint arXiv:2407.07530*, 2024.

Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Binhuraib, Antoine Bosselut, and Martin Schrimpf. From language to cognition: How llms outgrow the human language network. *arXiv preprint arXiv:2503.01830*, 2025.

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, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Frederico AC Azevedo, Ludmila RB Carvalho, Lea T Grinberg, José Marcelo Farfel, Renata EL Ferretti, Renata EP Leite, Wilson Jacob Filho, Roberto Lent, and Suzana Herculano-Houzel. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of comparative neurology*, 513(5):532–541, 2009.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Cordell M Baker, Joshua D Burks, Robert G Briggs, Andrew K Conner, Chad A Glenn, Justin M Robbins, John R Sheets, Goksel Sali, Tressie M McCoy, James D Battiste, et al. A connectomic atlas of the human cerebrum—chapter 5: The insula and opercular cortex. *Operative Neurosurgery*, 15(suppl_1):S175–S244, 2018a.

Cordell M Baker, Joshua D Burks, Robert G Briggs, Andrew K Conner, Chad A Glenn, Kathleen N Taylor, Goksel Sali, Tressie M McCoy, James D Battiste, Daniel L O'Donoghue, et al. A connectomic atlas of the human cerebrum—chapter 7: the lateral parietal lobe. *Operative Neurosurgery*, 15(suppl_1):S295–S349, 2018b.

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, et al. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46:e385, 2023.

Le Chang, Bernhard Egger, Thomas Vetter, and Doris Y Tsao. Explaining face representation in the primate brain using different computational models. *Current Biology*, 31(13):2785–2795, 2021.

Nathan Cloos, Moufan Li, Markus Siegel, Scott L Brincat, Earl K Miller, Guangyu Robert Yang, and Christopher J Cueva. Differentiable optimization of similarity scores between models and brains. *arXiv preprint arXiv:2407.07059*, 2024.

Jacob Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992. doi: 10.1037/0033-2909.112.1.155.

Colin Conwell, Jacob Prince, George Alvarez, and Talia Konkle. The unreasonable effectiveness of word models in predicting high-level visual cortex responses to natural images. In *Conference on computational cognitive neuroscience*, 2023.

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

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1):795–828, 2012.

Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz Kandola. On kernel-target alignment. *Advances in neural information processing systems*, 14, 2001.

Geoff Cumming and SUE Finch. A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and psychological measurement*, 61(4): 532–574, 2001.

Jörn Diedrichsen, Eva Berlot, Marieke Mur, Heiko H Schütt, Mahdiyar Shahbazi, and Nikolaus Kriegeskorte. Comparing representational geometries using whitened unbiased-distance-matrix similarity. *arXiv preprint arXiv:2007.02789*, 2020.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems*, 34:1556–1568, 2021.

Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lindsay, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24(7):431–450, 2023.

Samira M Epp, Gabriel Castrillón, Beijia Yuan, Jessica Andrews-Hanna, Christine Preibisch, and Valentin Riedl. Bold signal changes can oppose oxygen metabolism across the human cortex. *Nature neuroscience*, pp. 1–12, 2025.

Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392 (6676):598–601, 1998.

Dawn Finzi, Eshed Margalit, Kendrick Kay, Daniel LK Yamins, and Kalanit Grill-Spector. A single computational objective drives specialization of streams in visual cortex. *bioRxiv*, pp. 2023–12, 2023.

John Forrest, Ted Ralphs, Stefan Vigerske, Haroldo Gambini Santos, John Forrest, Lou Hafer, Bjarni Kristjansson, jpfasano, EdwinStraver, Jan-Willem, Miles Lubin, rlougee, a andre, jpgoncal1, Samuel Brito, h-i gassmann, Cristina, Matthew Saltzman, tosttost, Bruno Pitrus, Fumiaki MATSUSHIMA, Patrick Vossler, Ron @ SWGY, and to st. coin-or/cbc: Release releases/2.10.12, August 2024. URL `https://doi.org/10.5281/zenodo.13347261`.

Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in neural information processing systems*, 33: 13890–13902, 2020.

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.

Olivia Guest and Andrea E Martin. On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, 6(2):213–227, 2023.

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

Martin N Hebart and Chris I Baker. Deconstructing multivariate decoding for the study of brain function. *Neuroimage*, 180:4–18, 2018.

Larry V Hedges. Distribution theory for glass's estimator of effect size and related estimators. *journal of Educational Statistics*, 6(2):107–128, 1981.

Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, 12(1):6456, 2021.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Shunsuke Kamiya, Jun Kitazono, and Masafumi Oizumi. Koopman operator based dynamical similarity analysis for data-driven quantification of distance between dynamics. In *ICLR 2024 Workshop on Representational Alignment*, 2024.

Philipp Kaniuth and Martin N Hebart. Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior. *NeuroImage*, 257:119294, 2022.

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of neuroscience*, 17(11):4302–4311, 1997.

Chaitanya Kapoor, Sudhanshu Srivastava, and Meenakshi Khosla. Bridging critical gaps in convergent learning: How representational alignment evolves across layers, training, and distribution shifts. *arXiv preprint arXiv:2502.18710*, 2025.

Seyed-Mahdi Khaligh-Razavi, Linda Henriksson, Kendrick Kay, and Nikolaus Kriegeskorte. Fixed versus mixed rsa: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *Journal of Mathematical Psychology*, 76:184–197, 2017.

Meenakshi Khosla and Alex H Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, pp. 326–341. PMLR, 2024.

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.

Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1):491, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Nikolaus Kriegeskorte and Jörn Diedrichsen. Peeling the onion of brain representations. *Annual review of neuroscience*, 42(1):407–432, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N Apurva Ratan Murty, et al. Modeling short visual events through the bold moments video fmri dataset and metadata. *Nature communications*, 15(1):6241, 2024.

Baihan Lin and Nikolaus Kriegeskorte. The topology and geometry of neural representations. *Proceedings of the National Academy of Sciences*, 121(42):e2317881121, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

David Marr. *Vision*. W.H. Freeman and Company, 1982.

Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

Alex Murphy, Joel Zylberberg, and Alona Fyshe. Correcting biased centered kernel alignment measures in biological and artificial neural networks. *arXiv preprint arXiv:2405.01012*, 2024.

Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. Multimodal memorability: Modeling effects of semantics and decay on video memorability. In *European Conference on Computer Vision*, pp. 223–240. Springer, 2020.

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.

Mitchell Ostrow, Adam Eisen, Leo Kozachkov, and Ila Fiete. Beyond geometry: Comparing the temporal structure of computation in neural circuits with dynamical similarity analysis. *Advances in Neural Information Processing Systems*, 36:33824–33837, 2023.

Fernando M Ramírez. Representational confusion: the plausible consequence of demeaning your data. *bioRxiv*, pp. 195271, 2017.

Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

Edmund T Rolls, Gustavo Deco, Chu-Chung Huang, and Jianfeng Feng. The human language effective connectome. *NeuroImage*, 258:119352, 2022.

Edmund T Rolls, Gustavo Deco, Chu-Chung Huang, and Jianfeng Feng. Multiple cortical visual streams in humans. *Cerebral Cortex*, 33(7):3319–3349, 2023.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Maria Ryskina, Greta Tuckute, Alexander Fung, Ashley Malkin, and Evelina Fedorenko. Language models align with brain regions that represent concepts across modalities. *arXiv preprint arXiv:2508.11536*, 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.

Rebecca Saxe, Matthew Brett, and Nancy Kanwisher. Divide and conquer: a defense of functional localizers. *Neuroimage*, 30(4):1088–1096, 2006.

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.

Shaan Shah and Meenakshi Khosla. Representational alignment across model layers and brain regions with hierarchical optimal transport. *arXiv preprint arXiv:2510.01706*, 2025.

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, 2024.

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

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Yingtian Tang, Abdulkadir Gokce, Khaled Jedoui Al-Karkari, Daniel Yamins, and Martin Schrimpf. Diverse perceptual representations across visual pathways emerge from a single objective. *bioRxiv*, pp. 2025–07, 2025.

Imran Thobani, Javier Sagastuy-Brena, Aran Nayebi, Jacob S Prince, Rosa Cao, and Daniel LK Yamins. Model-brain comparison using inter-animal transforms. In *8th Annual Conference on Cognitive Computational Neuroscience*, 2025.

William Timkey and Marten Van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, 2021.

David C Van Essen, Matthew F Glasser, Donna L Dierker, John Harwell, and Timothy Coalson. Parcellations and hemispheric asymmetries of human cerebral cortex analyzed on surface-based atlases. *Cerebral cortex*, 22(10):2241–2262, 2012.

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.

Bin Wang, Jiayue Guo, Tianyi Yan, Seiichiro Ohno, Susumu Kanazawa, Qiang Huang, and Jinglong Wu. Neural responses to central and peripheral objects in the lateral occipital cortex. *Frontiers in human neuroscience*, 10:54, 2016.

Liang Wang, Ryan EB Mruczek, Michael J Arcaro, and Sabine Kastner. Probabilistic maps of visual topography in human cortex. *Cerebral cortex*, 25(10):3911–3931, 2015.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in neural information processing systems*, 34:4738–4750, 2021.

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.

REGION MATCHING AND FEATURE SPECIFICITY: SHORT RECOMMENDATIONS

Box 1: Why region matching?

- It sanity-checks that the benchmark is sensitive to what distinguishes a target region. If it fails, conclusions about a region are limited.
- It catches preprocessing and measure choices that render a region indistinguishable.
- It can serve as diagnostic tool to investigate what measures are (in)sensitive to.

Implementing region matching

Implementation choices should match or include those intended for use in the benchmark. We recommend not using the benchmark's test split at all.

Aspect	Requirements	This work
Distractor regions	Regions that the benchmark is expected to tell apart from the target. Motivate exclusions of other benchmarked regions.	Remaining analyzed regions, which exclude direct neighbor regions.
Region definition	Matching what is used in the target benchmark - e.g., anatomical, or functional.	Anatomical boundaries of Glasser regions, based on group surface maps.
Source $\tilde{X}_r$, derived from data of region r	Should represent a 'good model' of that region. Can be another subjects region data $X_r^{s_j}$, or where appropriate, subject-aggregate data.	Leave-one-subject-out averaged fMRI map, leaving out subject s_i to generate sources $\tilde{X}_r$ for target subject s_i.
Quantification	Optimally invariance to monotonous transformations of the measure.	Accuracy; Cohen's d_z (not fully invariant)
Identifiability threshold	Conservative interpretation $\Longrightarrow$ accuracy threshold = minimal tolerable accuracy on an ideal model zoo. Alternatively, estimate for correctness of region matching on a second instantiation of the benchmark, when using the benchmark's evaluation procedure. Require at least $d_z > 0$ or correct majority match.	For H.3, we vary the threshold.

Acting on region matching results

If region matching fails for a region, modify the benchmark to make region matching pass for all its target regions. For example:

- Regions: Drop the failing region, merge regions that are often confused with each other, try a coarser set of regions.
- Vary preprocessing.
- Similarity measure: If another measure exists that is considered equally justified, and which shows better identifiability across target regions, a measure change can be considered.

Box 2: Feature specificity

- Feature specificity is a measure property, our test catches lack of it on benchmark data.
- We combine two directions into one score; both tested empirically:
  a) Is the measure sensitive to superfluous feature dimensions present?
  b) Is it sensitive to a subset of features missing?

  If a) fails, a high-similarity match might just *contain* matching features. b) is generally useful.

Testing for lack of feature specificity

We recommend testing with data from the modalities used in the benchmark.

Aspect	Requirements	This work
Feature map r to append to	r contains features that the measure is sensitive to	We require $min(d_z) > 0$, i.e., the closest distractor region having $d_z > 0$.
Confounder features added	(i) Not a better match to the target than the features they are added to. (ii) Clearly mismatching - associated with other functionality than r.	Large set of spatially separated regions
Dropped features	Associated with partially different functionality than the remaining features	We drop a large part of the compound region, including many functionally different sub-regions.
Threshold	Allow slack for noise-related mismatches. Require accuracy well above the chance level of 50%, separately for detecting dropped and added features.	CI excluding chance; higher is better.

When to require feature specificity, and implications when it is lacking

- Feature specificity is a prerequisite for claims about similarity of the mechanism implemented by both systems.
- Awareness of the lack of feature specificity helps to avoid wrong conclusions. A non-feature specific measure might still be a helpful first indicator.
- A higher score under a non-feature-specific measure might reflect an unspecific, larger breadth of learned features.
- When data covers only part of a region, a researcher might consciously choose a non-feature-specific measure, accepting weaker claims.

CONSIDERATIONS FOR IMPLEMENTING FEATURE SPECIFICITY TESTS

To evaluate feature specificity empirically, as described in section 3.2, we propose to use composite feature maps and test whether they can be distinguished from their components. In this process, if we were to append or drop, noise features, we would be testing noise robustness instead, and if we append features that closely match the original ones, chance results are not indicative of a lack of feature specificity. Therefore, we propose the following

Requirements for appended features: When comparing the similarity of region X_1 across subjects to the similarity between X_1 and a compound region, $R = [X_1, X_2]$ then the appended features X_2 must not contain features that are a better match to the target than the ones we add them to. Otherwise, match quality is in trade-off with a penalty for added mismatching features. Further, the added features X_2 must contain clearly mismatching features. This means that the compound region $R = [X_1, X_2]$ should *include* features that are associated with other functionality than X_1 .

Additionally we propose **requirements for dropped features:** For detection of missing features, we compare the similarity of region $R = [X_1, X_2]$ to itself across subjects, $S(\tilde{R}, R^{s_i})$, to the similarity between target region $R = [X_1, X_2]$ to source region X_1, $S(\tilde{X_1}, R^{s_i})$. Then, the dropped part X_2 should not be redundant to X_1, else we would not expect similarity to matching $[X_1', X_2']$ to decrease when dropping X_2. For example, in fMRI data we would not want to remove every second voxel since voxels are spatially correlated. Instead, we similarly require that the dropped part X_2 include features associated with different functionality than X_1, for a symmetric argument: beyond guaranteeing non-redundancy, this ensures we are not just evaluating the effect of noise removal, if X_2 was noise or a mix of redundant features to X_1 and noise.

While on brain data, we can use the benchmark's assumptions which regions are functionally distinct, to argue that added or dropped features fulfill these requirements. In DNNs, in contrast, this is harder to ensure, because DNNs' random initialization creates random features that might survive training even if they are not functionally used by the network. That is, it is harder to ensure that one set of DNN features does not contain a good match for another set of DNN features.

SIMILARITY MEASURES

All measures except RSA and CKA, which do not require fitting a model, use five-fold crossvalidation, fitting on each train split and evaluating on the evaluation split, then averaging the score over splits.

Soft matching Soft Matching (Khosla & Williams, 2024) is a 'softened' version of one-on-one unit matching, that supports unequal source and target feature map sizes by allowing splitting up of the units to match. It finds the optimal transportation polytope P that minimizes $\sum_{i,j} P_{ij} ||\mathbf{x_i} - \mathbf{y_j}||^2$, where $\mathbf{x_i}$ is the ith column of X, and equivalently for Y. Here, P is an $\mathbb{R}^{N_x \times N_y}$ matrix whose elements must sum to $1/N_y$ in each column, and to $1/N_x$ in each row, and $P_{ij} \leq 0 \ \ \forall \ i, j$. When $N_x = N_y$ this is equivalent to minimizing $||X - YP||_F^2$, where $||.||_F^2$ is the Frobenius norm. See Khosla & Williams (2024) for a detailed description and interpretation.

Pairwise matching A simple baseline, Pairwise Matching is an asymmetric method that matches each feature dimension (each voxel) in the target feature map Y, to the source dimension in X that it is most strongly correlated with. Dimensions of X can be used multiple times, leading to a possible bias towards assigning higher scores for larger source feature maps X.

Procrustes (Angular orthogonal Procrustes similarity) Also refered to as Procrustes shape metric (Williams et al., 2021), it is an old method that gained traction recently (Ding et al., 2021). To compute orthogonal Procrustes distance between feature maps X and Y, after centering along columns, we compute the optimal orthogonal transform R so that $||X - YR||_F^2$ is minimized, after which $S_{Procrustes}(X, Y) = -arccos(trace(X^T YR)/(||X||_F ||Y||_F))$, where $||.||_F$ is the Frobenius norm. Since $||A||_F^2 = trace(A^T A)$, the trace expression is proportional to the middle term of $||X - YR||_F^2$ after normalizing X and Y. Note that we multiply by -1 here to convert a distance into a similarity. Since our methods are invariant to shared shifts and linear scaling of all similarity values, we do not need to convert the range further.

CKA Centered Kernel Analysis (Kornblith et al., 2019; Cortes et al., 2012; Cristianini et al., 2001) (CKA) computes a normalized Hilbert-Schmidt Independence Criterion (HSIC), based on kernel matrices K_X, K_Y, where

$$(K_X)_{i,j} = k(X_{i,:}, X_{j,:})$$

is the kernel function k applied between the vector of voxel activations of stimulus i and the one for stimulus j . CKA computed between two feature maps X and Y is:

$$S_{CKA}(X, Y) = \frac{HSIC(K_X, K_Y)}{\sqrt{HSIC(K_X, K_X) \cdot HSIC(K_Y, K_Y)}}$$

with $HSIC(K, L) = \frac{1}{(n-1)^2} tr(KHLH)$, with H a centering matrix. We use linear CKA where K is based on a linear kernel $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$, applied to X and Y after centering over columns (i.e., voxels).

RSA variants Representational Similarity Analysis (RSA, (Kriegeskorte et al., 2008)) is, like CKA, a second-order method that compares distance patterns between stimuli. First, Representational Dissimilarity Matrices (RDMs) encode at their (i, j)th element the distance $d(\mathbf{x}_i, \mathbf{x}_j)$ between voxel activities for stimulus i and stimulus j, i.e. x_i is the ith row of a $(N_{stimuli} \times N_{features})$ feature map X. d can be any distance function; we here include the Pearson correlation-based distance,

$1 - corr(\mathbf{x}_i, \mathbf{x}_j)$ (we denote this version just 'RSA'), cosine distance $1 - \frac{\mathbf{x}_i^T \cdot \mathbf{x}_j}{||\mathbf{x}_i|| \cdot ||\mathbf{x}_j||}$, Euclidean distance $||\mathbf{x}_i - \mathbf{x}_j||_2$, univariate distance which we define as the absolute difference of mean region activity per stimulus, i.e., $|\bar{x}_i - \bar{x}_j|$ for stimuli i and j. And finally crossvalidated Euclidean squared distance (Walther et al., 2016), $(\mathbf{x'}_i - \mathbf{x'}_j)^T \cdot (\mathbf{x''}_i - \mathbf{x''}_j)$, where unlike in all other measures, we average over two, not three trials to obtain $\mathbf{X}'$ and take the dot product to the activations of the remaining trial, $\mathbf{X}''$, averaging results across the three possible splits. When combined with whitening via a shrunk covariance (section 3.3), we denote the resulting measure 'Crossnobis' in reference to (Walther et al., 2016), although our covariance computation differs from theirs. In both RSA and veRSA we then report Kendall's rank correlation of the upper-triangular parts of the RDMs computed for the two featuremaps to compare.

veRSA variants Voxel-wise-encoding RSA (veRSA) (Khaligh-Razavi et al., 2017; Konkle & Alvarez, 2022) first fits a linear model to predict voxel activations of the target features based on the source features. It then computes RSA between the predicted and true voxel activations on a held-out set of stimuli. In the standard variant, referred to just as 'VERSA', we use Pearson correlation distance to create RDMs. We also study Euclidean and Univariate versions, using the corresponding distance functions to create RDMs.

Linear predictivity Linear predictivity is an asymmetric measure fitting a linear model to ($N_{stimuli} \times d$) source features X to fit ($N_{stimuli} \times d'$) target features Y, and evaluating mean Pearson correlation between predicted and true target features:

$$S_{LP}(\mathbf{X}, \mathbf{Y}) = \frac{1}{d'} \sum_{j=1}^{d'} corr(\mathbf{X}\mathbf{a_j} + b_j, \mathbf{Y}\mathbf{e_j})$$

Where e_j is the j-th unit vector. We compute linear model fits using RidgeCV from the scikit-learn python package. It determines an L^2 regularization weight from a range (we use $[10^{-8}, 10^{-7}, ..., 10^{8}]$) based on leave-one-out crossvalidation. We apply five-fold crossvalidation on top.

ReverseLinearPredictivity is exactly LinearPredictivity with the roles of source and target feature maps reversed: $S_{RevLP}(\mathbf{X}, \mathbf{Y}) = S_{LP}(\mathbf{Y}, \mathbf{X})$. It is commonly not used as standalone measure, but as additional measure to S_{LP} (Higgins et al., 2021; Chang et al., 2021).

SymmetricLinearPredictivity is the mean of reverse and standard linear predictivity: $S_{SymLP} = \frac{1}{2}(S_{LP}(\mathbf{X}, \mathbf{Y}) + S_{RevLP}(\mathbf{X}, \mathbf{Y}))$. Unlike S_{LP} and S_{RevLP} it is symmetric and is not directly biased towards source or target feature maps with larger effective dimensionality.

PLS regression Partial least squares regression is the measure used in the first brain-score version (Schrimpf et al., 2018). We use class *PLSRegression* of scikit-learn. It iteratively finds the most strongly correlated components in X and Y and regresses Y^k on each component of X, where $Y^1 = Y$ and Y^{k+1} is the residual from the regression at step k.

DATASETS

NATURAL SCENES DATASET

The Natural Scenes Dataset (NSD), (Allen et al., 2022) consists of high-resolution (7T), whole-brain fMRI scans of eight human participants. Each participant viewed up to $10,000$ images of natural scenes taken from the COCO dataset (Lin et al., 2014). Of those, $1,000$ images were seen by all participants. Since each stimulus was shown thrice, our subset of NSD consists of $24,000$ trials. During the recording sessions, participants performed a continuous recognition task, i.e. they reported for each image whether they had seen it before or not. In each trial, an image was presented for three seconds.

BOLD MOMENTS DATASET

The BOLD Moments Dataset (BMD) consists of medium-resolution (3T), whole-brain fMRI scans of ten human participants. The stimuli consist of $1,102$ video clips sampled from the Memento10k video memorability dataset (Newman et al., 2020), each three seconds long. We used the train split of the dataset, consisting of $1,000$ videos repeated thrice. We dropped samples corresponding to video IDs 258 and 831, for reasons unrelated to this project, for a total of $2,994$ trials. The clips capture ecologically meaningful scenes from everyday life, showing a diverse range of real-world actions and

events. Participants were tasked to report null trials (in which no video was shown) with a button press.

NOISE CEILING SIGNAL-TO-NOISE RATIO (NCSNR)

To estimate the noise level in a voxel, we use the NCSNR, following (Allen et al., 2022) and (Lahner et al., 2024). We use the NCSNR values provided by both datasets. NCSNR is based on the assumption that noise is additive, has zero-mean and is normally distributed, and considers all signal variation across repeats of the same stimulus as noise (Allen et al., 2022). After z-scoring the beta values per voxel and per session, noise and signal standard deviation are computed as follows, for each voxel:

$$\hat{\sigma}_{noise} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_{rep,i}-1}\sum_{j=1}^{3}(\beta_j^i - \bar{\beta}_i)^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}^2_{\beta_i}},$$

where $N_{rep,i}$ is the number of repeats for stimulus i, N is the number of stimuli, $\bar{\beta}_i$ is the mean β-value for this voxel and stimulus i. $\hat{\sigma}^2_{\beta_i}$) is the empirical variance of β values for this voxel and stimulus i.

Assuming each voxels variance is 1 because of z-scoring, using the law of total variance to decompose 1 into $\hat{\sigma}^2_{signal}$, which is the variance of stimulus means, plus $\hat{\sigma}^2_{noise}$, the mean of per-stimulus variances, signal standard deviation $\hat{\sigma}_{signal}$ is then computed as

$$\hat{\sigma}_{signal} = \sqrt{|1 - \hat{\sigma}^2_{noise}|_+}$$

where $|.|_+$ caps values to be ≥ 0, to prevent errors from small inaccuracies driving values below zero. Finally, the noise ceiling signal-to-noise ratio (NCSNR) is

$$ncsnr = \frac{\hat{\sigma}_{signal}}{\hat{\sigma}_{noise}},$$

i.e., higher values represent a higher SNR.

NCSNR DOES NOT PREDICT IDENTIFIABILITY RELIABLY

Fig. 6 shows the relation of median NCSNR within a region - measured before dropping voxels below the NCSNR threshold - to a regions identifiability. We quantify a regions 'overall' identifiability by averaging the accuracies of the two best-performing measures for the region (top panels), and separately per measure (bottom panel). While correlated, it is not possible to apply a threshold on the median NCSNR, to separate low-accuracy from high-accuracy regions - this will either drop high-accuracy regions or include low-accuracy regions. That is, using standard noise metrics to inform which regions to use in a benchmark does not appear appropriate, in particular since usually, a single measure would be used in a benchmark.

PREPROCESSING AND VOXEL SELECTION METHODS

PREPROCESSING: WHITENING

We compute a pooled noise covariance matrix by averaging over per-stimulus covariances, each computed across trials. For NSD we base the computation on data of the ca. 700 stimuli that are shared across subjects and have three trials available across all subjects, for BMD we use the same stimuli as in the remaining experiments. Since shrinkage methods such as Ledoit-Wolf used in (Walther et al., 2016) do not translate directly to pooled covariance estimation, we follow the recommendation in (Diedrichsen et al., 2020) of choosing a value between 0.2 and 0.4 and use a fixed shrinkage factor of $h = 0.3$, so that when $\hat{\Sigma}$ is the pooled empirical covariance, the shrunk covariance is $\Sigma = h \cdot diag(\hat{\Sigma}) + (1-h)\hat{\Sigma}$.

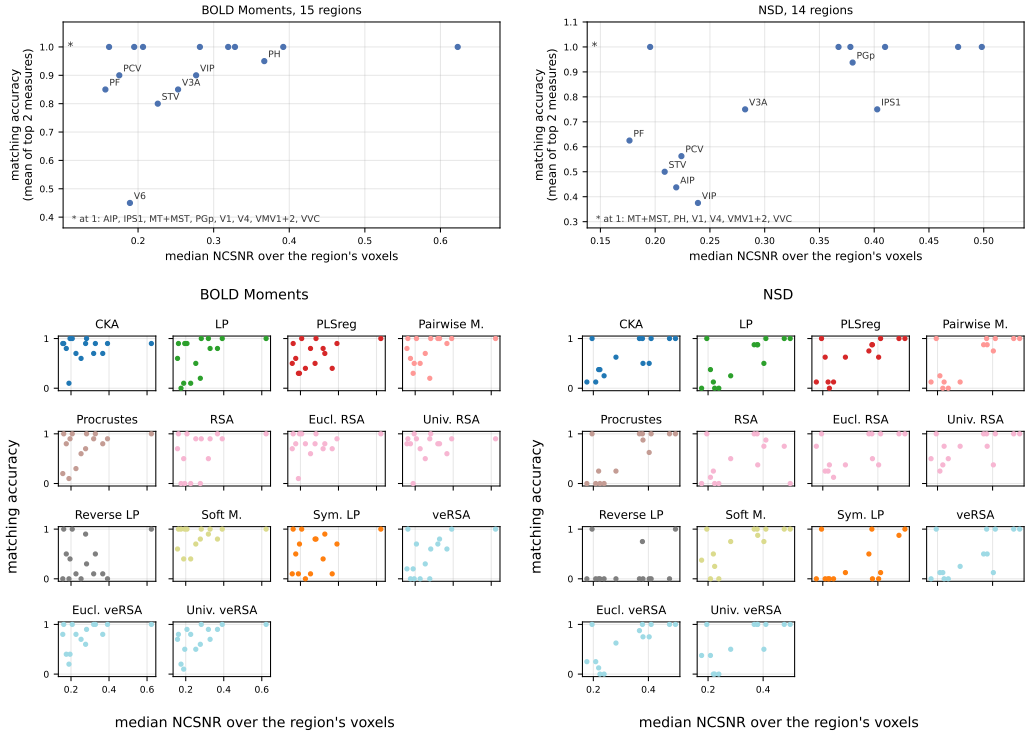


Figure 6: **NCSNR vs region matching accuracy,** aggregate (top) and per measure (bottom). **Top:** We estimate a region's identifiability as the mean accuracy of the two measures that score highest accuracy for the region. This is shown on the y-axis, compared to median NCSNR on the y axis. **Bottom:** Matching accuracy per measure, compared to median NCSNR. One dot represents one region in both plot types. Interpretation in section E.1. **Note that x-axis scales differ between BMD and NSD plots.**

VOXEL SELECTION METHODS

As standard voxel selection, we include only voxels above a noise threshold of 0.2 NCSNR. As second variant, we include all voxels. To test region matching on equal-sized feature maps, a third option consists of picking those 50 voxels with highest NCSNR per region, and excluding subjects where a region has fewer than 50 voxels above the threshold of 0.2 NCSNR. The ablation is shown in figure 15.

REGION MATCHING: RECOMMENDATIONS AND ADDITIONAL DETAILS

Generally, for sanity-checking a similarity benchmark, the setup should reflect the desired benchmark setting as closely as possible.

To avoid overfitting through measure selection, we recommend excluding the benchmark test set from sanity check computations. We do this for BMD but not for NSD, since in the latter there is no standard train/test split that takes stimulus identity into account. We note that train/test splits with shared subjects mean that choosing a measure based on benchmark results on the train set, can introduce subject-specific selection effects into the results. However, since DNN-to-brain comparisons are only part of the downstream benchmark, direct overfitting on the test set is not a risk.

QUANTIFY REGION MATCHING PERFORMANCE: ACCURACY AND COHEN'S d_z

We generally propose that region matching is best used with pairwise statistical methods, where region self-similarity is compared to similarities to other regions *for the same target subject*. Pairwise comparisons are important when there are strong subject-specific effects, as in NSD (section K.3).

Formally, we assume we have a candidate similarity measure S and a dataset of brain recordings from multiple regions and multiple human observers in response to a shared set of N stimuli. We write $X_j^{s_i}$ for the $(N \times d_{s_i,j})$ matrix of features obtained from subject s_i across region j. We denote assumed 'ground truth models' for region j and subject s_i as $\tilde{X}_j^{s_i}$. (In our case those are leave-subject-s_i-out averaged feature maps $\tilde{X}_j^{s_i} = \overline{X}^{-s_i}$).

REGION MATCHING ACCURACY

Hiding subject indices, we then want to test whether

$$S(M_j, X_j) > S(M_k, X_j) \quad \forall\, k \in \text{distractors}(j) \tag{2}$$

The condition (eq. (2)) translates to a binary accuracy value per subject and target region:

$$\delta_{s_i,j} := \begin{cases} 1 & \text{if } S(\tilde{X}_j^{s_i}, X_j^{s_i}) > S(\tilde{X}_k^{s_i}, X_j^{s_i}), \\ & \quad \forall\, k \in \text{distractors}(j) \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

For example, to check whether a measure can differentiate region V1 from regions V4 and MT in subject s_1, we calculate the similarity values between 'models' of V1, V4 and MT to subject s_1's brain activity for V1. Now $\delta_{s_i,V1} = 1$ if the measure assigns the highest similarity of the three to the V1-'candidate model' ($\overline{X}_{V1}^{-s_1}$), and $\delta_{s_i,V1} = 0$ if either the V4 or the MT 'model' was assigned a higher similarity.

We then report the overall mean accuracy per measure:

$$\text{accuracy}(S) := \frac{1}{N_{\text{subj}}} \sum_{i=1}^{N_{\text{subj}}} \frac{1}{N_{\text{roi}}} \sum_{j=1}^{N_{\text{roi}}} \delta_{s_i,j} \tag{4}$$

computing bootstrap confidence intervals over subjects. We use a fixed set of target regions (see section G.3), treating the remaining regions from the set as distractors. The overall mean accuracy is specific to the chosen set of regions.

COHEN'S d_z

In addition, we consider Cohen's d_z (Cohen, 1992), a measure of effect size. It is a more nuanced metric, taking into account how much below or above self-similarity is compared to the similarity to the strongest distractor. The two metrics broadly share a decision boundary, see fig. 47 and fig. 46.

It is computed between two paired sets of values, as the standardized mean of the differences between each pair:

$$d_z = \frac{\bar{\Delta}}{\hat{\sigma}_\Delta}, \tag{5}$$

where

$$\hat{\sigma}_\Delta = \sqrt{\frac{\sum_{i=1}^{n}(\Delta_i - \bar{\Delta})^2}{n-1}}, \tag{6}$$

Δ_i is the ith paired difference and $\bar{\Delta} = \frac{\sum_{i=1}^{n}\Delta_i}{n}$ is their mean. In our case, for a target ROI j, the first set of values are the $n = N_{subj}$ self-similarity scores, one per subject s. Those are the similarities between the 'models' $\tilde{X}_j^s$ and the target ROI feature map X_j^s. The second set are the similarity values between the 'wrong' models $\tilde{X}_k^s$ for a single distractor region $k \neq j$, to the target regions' features X_j^s. That is, we have one d_z score per distractor region $k \neq j$, computed from paired differences

$$\Delta^i = \Delta_{j;k}^i = S(\tilde{X}_j^{s_i}, X_j^{s_i}) - S(\tilde{X}_k^{s_i}, X_j^{s_i}). \tag{7}$$

We report *minimum* Cohen's d_z across distractor regions, as a measure of the sensitivity of the measure s, for target ROI j. This way, if at least one of the other ROIs cannot be distinguished from the target, the overall score is low.

CHOICE OF 'GROUND TRUTH MODEL'

Generally, when choosing how to convert other subjects' neural data into a 'ground truth model' to compare against, the goal is to choose whatever we expect to be the closest representation to the one we test.

We use the *averaged feature maps* from the remaining subjects as 'models' to reduce the impact of subject-specific anatomic differences, which is also reflected in higher similarity values, see fig. 48 and fig. 49, and higher noise ceiling signal-to-noise ratio (NCSNR) (fig. 12).

With other modalities or other region-selection methods, choices such as other subjects' individual feature maps will be more appropriate. In particular in settings with zero unit-to-unit correspondence such as electrophysiological recordings or with regions determined by functional localizers.

REGIONS OF INTEREST

Exact anatomic location of a (functional) region can vary substantially between subjects (Wang et al., 2015; Glasser et al., 2016; Wang et al., 2016), which can lead to false (mis-) matches and obscure results. Therefore, we exclude adjacent regions, leading to a checkerboard-like pattern of 15 (14) selected regions in BMD (NSD) (fig. 7 and table 1). See details in section G.5 and section G.6.

PREPROCESSING ABLATION

Unless otherwise stated, for each ROI we select the set of (subject-specific) voxels of at least an NCSNR of 0.2 to remove noisy voxels, as is also done in (Conwell et al., 2024). We ablate this as part of

ROI	Stream	Description
V1	both (early)	Primary visual cortex: first cortical stage of visual input processing, retinotopic mapping
V2	both (early)	Secondary visual cortex: processes orientation, depth, texture, integrates from V1
V4	both (early)	Early-to-intermediate visual area: involved in texture, color, form, object features
VVC	ventral	Ventral Visual Complex (inferior occipito-temporal) – higher order object/face/scene processing
VMV1, VMV2	ventral (medial)	Ventromedial visual area 1 and 2: part of a ventromedial visual pathway implicated in spatial scene processing
V6	dorsal	Dorsal stream occipital region: self-motion, peripheral vision
IPS1	dorsal	Intraparietal sulcus area 1: retinotopic mapping in parietal cortex, visuospatial attention
V3A	dorsal	Visual area V3A: motion, complex stimuli, transition between dorsal stream and early visual areas
MT	dorsal	Middle temporal area (also called V5): motion perception, speed, direction of visual motion
MST	dorsal	Medial superior temporal area: more complex motion patterns (expansion, rotation), optic flow
PH	dorsal	Part of MT+ complex, motion
PGp	other (parietal)	Posterior inferior parietal lobule region PGp: higher-order visuospatial and associative region
STV	other (temporal)	Superior temporal visual region – language-implicated (Rolls et al., 2023) region, also associated with higher-level cognitive functions(Baker et al., 2018a) . Possibly involved in social/facial motion processing (Rolls et al., 2022).
AIP	other (parietal)	Anterior intraparietal area: grasping/reach-related visuomotor processing (Baker et al., 2018b)
VIP	other (parietal)	Ventral intraparietal area: region integrating visual and somatosensory information
PF	other (parietal)	Parietal area PF (inferior parietal lobule) – higher-order multimodal region, possibly involved in tool usage and body image representation
PCV	other (parietal)	Precuneus Visual Area. Visuo-motor area (Rolls et al., 2022).

Table 1: **Summary of ROIs included in our analyses.** We state each ROI's affiliation to a stream, and associated functional roles.

Anatomical reliability based on Glasser-atlas probabilistic map

Glasser et al (Glasser et al., 2016) created a brain parcellation based on multiple types of MRI data for 210 young adults (ages 22 to 35) from the Human Connectome Project, valid on a similar second set of 210 subjects. Parcellations were created for each individual subject in a semi-automated fashion. Based on those individual parcellations, each voxel in a standardized space can be assigned a probability of belonging to each of the 180 regions, as the fraction of participants in which the voxel was assigned to the region. This yields a probabilistic map, with a region-probability distribution for each voxel. A group parcellation is then created by assigning each voxel to the region for which it has highest probability (i.e., the region to which it was assigned in the largest number of participants.) Validation showed high reliability of the resulting group parcellation (Glasser et al., 2016). Since the datasets BMD and NSD do not provide subject-specific parcellations for most regions, our analyses are based on a group-average parcellation (Glasser's parcellation). Using group-based parcellations is a common practice in brain-DNN similarity benchmarking (Sartzetaki et al., 2024; Tang et al., 2025). We use the Glasser atlas probabilistic map based on 210 validation subjects, with a slightly more balanced male-female participant ratio.

ROI	Min. no voxels included	Median no voxels included	Max. no voxels included	Mean Glasser-atlas probability
V1	545	594	993	0.93
V4	406	453	473	0.72
VVC	155	186	202	0.75
VMV1+2	95	162	234	0.75
IPS1	146	187	199	0.77
V3A	56	155	207	0.82
MT+MST	188	205	222	0.80
PH	233	278	286	0.81
PGp	159	238	251	0.84
STV	63	131	188	0.73
AIP	184	315	488	0.82
VIP	50	136	201	0.73
PF	31	207	506	0.87
PCV	104	138	194	0.78
V6	2	24	50	0.75

Table 2: Number of voxels surviving NCSNR thresholding, in NSD. Shown are minimum, median and maximum across subjects, per ROI. Also shown is the mean Glasser-atlas based region probability within each region, see also fig. 10.

ROI	Min. no voxels included	Median no voxels included	Max. no voxels included	Mean Glasser-atlas probability
V1	361	556	1221	0.93
V4	451	589	630	0.72
VVC	151	179	209	0.75
VMV1+2	90	164	236	0.75
IPS1	123	215	243	0.77
V3A	55	225	273	0.82
MT+MST	221	244	249	0.80
PH	258	333	371	0.81
PGp	168	219	294	0.84
STV	75	151	219	0.73
AIP	121	267	563	0.82
VIP	91	182	203	0.73
PF	92	236	580	0.87
PCV	50	112	264	0.78
V6	11	104	198	0.75

Table 3: Same as table 2 but for BMD, and the regions we use for analyses on BMD.

Region selection based on anatomical reliability, overlap and on NCSNR

We selected the regions used in our main analyses as following, based on the BMD dataset. For NSD we drop V6, so that all remaining regions fulfill the same criteria on NSD:

1. **Select a set of stimulus-responsive regions as candidate ROIs.** We select all regions in the respective dataset that both i) have a median voxelwise NCSNR of at least 0.15 and ii) in each subject, at least 10 voxels are above our voxel-inclusion NCSNR threshold of 0.2. The first criterion excludes regions containing mostly noisy or non-responsive voxels, the second criterion ensures a minimum amount of 10 features in all feature maps involved in our computations. (The minimum and median number of voxels surpassing the threshold is shown in tables 3 and 2.)
2. **Select a checkerboard-pattern subset of regions.** For each dataset, we manually select a subset of regions that are non-overlapping, as defined by the probabilistic map (validation set) of the Glasser atlas (Glasser et al., 2016): two regions are considered non–overlapping if the mean probability of the voxels in region one, of belonging to region 2, does not exceed

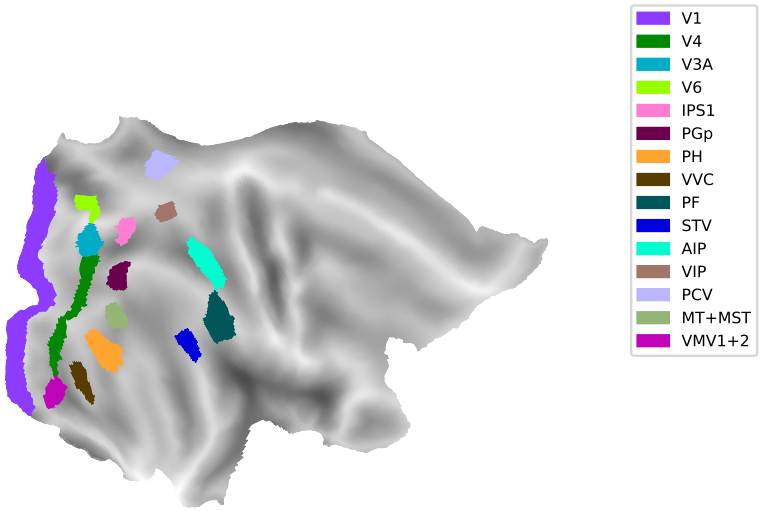


Figure 7: The checkerboard-pattern-subset of regions used in all our analyses on BMD, except if stated otherwise. For NSD, we drop region V6 because too few of its voxels survive NCSNR thresholding.

3 % (0.03), and the reverse. Asymmetric overlaps defined this way are shown in fig. 9. We also require a minimum anatomical reliability, as measured by the mean Glasser-atlas probability across voxels within a region, of belonging to this region, of 0.72. This threshold was chosen to still be able to include region V4. Mean Glasser-atlas probabilities for selected and all candidate regions are shown in fig. 10.

The final selected regions are shown in figure 7 (BMD).

Low SNR and low voxel number alone cannot explain low d_z values of regions STV, AIP, VIP, PF, PCV in NSD; potentially explains low d_z of V6 in BMD.

As shown in main results figure 2, identifiability of a range of regions seems much worse in NSD compared to BMD. Is this reflected in lower NCSNR values in those regions in NSD, or is it caused by a small number of voxels surviving thresholding, thus possibly leading to unreliable results? Figures 11 shows the distribution of NCSNR values across regions, for both datasets (see also fig. 8, bottom). The regions in question are on the right of both plots, next to V6. In STV, median NCSNR is lower in NSD compared to BMD, in PCV we observe the opposite; no clear trend across all affected regions is visible.

Tables 3 and 2 show how many voxels survived NCSNR thresholding per region, in the two datasets. For all regions, the median included number of voxels is above 100 in both regions. **For V6, the smallest number of included voxels is small (11)**, which could be part of the reason for its low values of d_z across measures in BMD. In NSD, PF with only 31 voxels included in one subject might also be affected by this. However, the other regions affected (STV, AIP, VIP, PCV) include a minimum of 50 voxels and above, PCV even having twice the minimum number of voxels in NSD, compared to BMD. Neither NCSNR differences nor small numbers of included voxels therefore seems a likely explanation of the differences between d_z values between datasets.

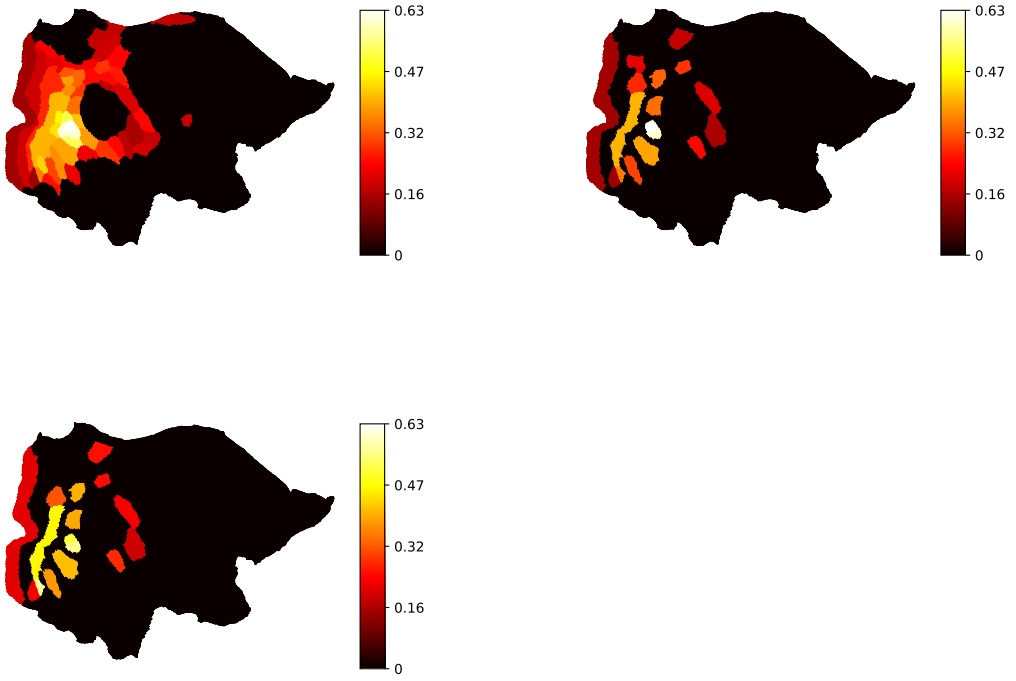


Figure 8: **Top:** Glasser-atlas regions (colored by median NCSNR) exceeding a median NCSNR of 0.15 and with at least 10 voxels exceeding our NCSNR threshold of 0.2, for our default set of regions for BMD (see fig. 7). **Center row:** Voxel-median subject-mean NCSNR per region for BMD, for the same set of regions. **Bottom row:** Same for NSD (dropping region V6 which we exclude for NSD experiments).

Additional Results

Region matching and discriminability

See fig. 13.

Voxel selection ablation

See fig. 15.

Preprocessing ablation

See fig. 16, fig. 17 and fig. 18.

RSA and veRSA variants

Figure 19 shows veRSA variants on BMD and NSD, and RSA variants on NSD, complementary to fig. 4 – except we do not test whitening and crossvalidation variants for veRSA (see preprocessing ablation section for whitening results). We see a similar pattern for both RSA and veRSA, where Pearson-correlation based RDMs lead to reduced region matching accuracy. Further, fig. 20 ablates individual attributes of our 'Crossnobis' implementation - which is crossvalidated Euclidean RSA with whitening with shrinkage 0.3, with no z-scoring beforehand and with centering. And fig. 21

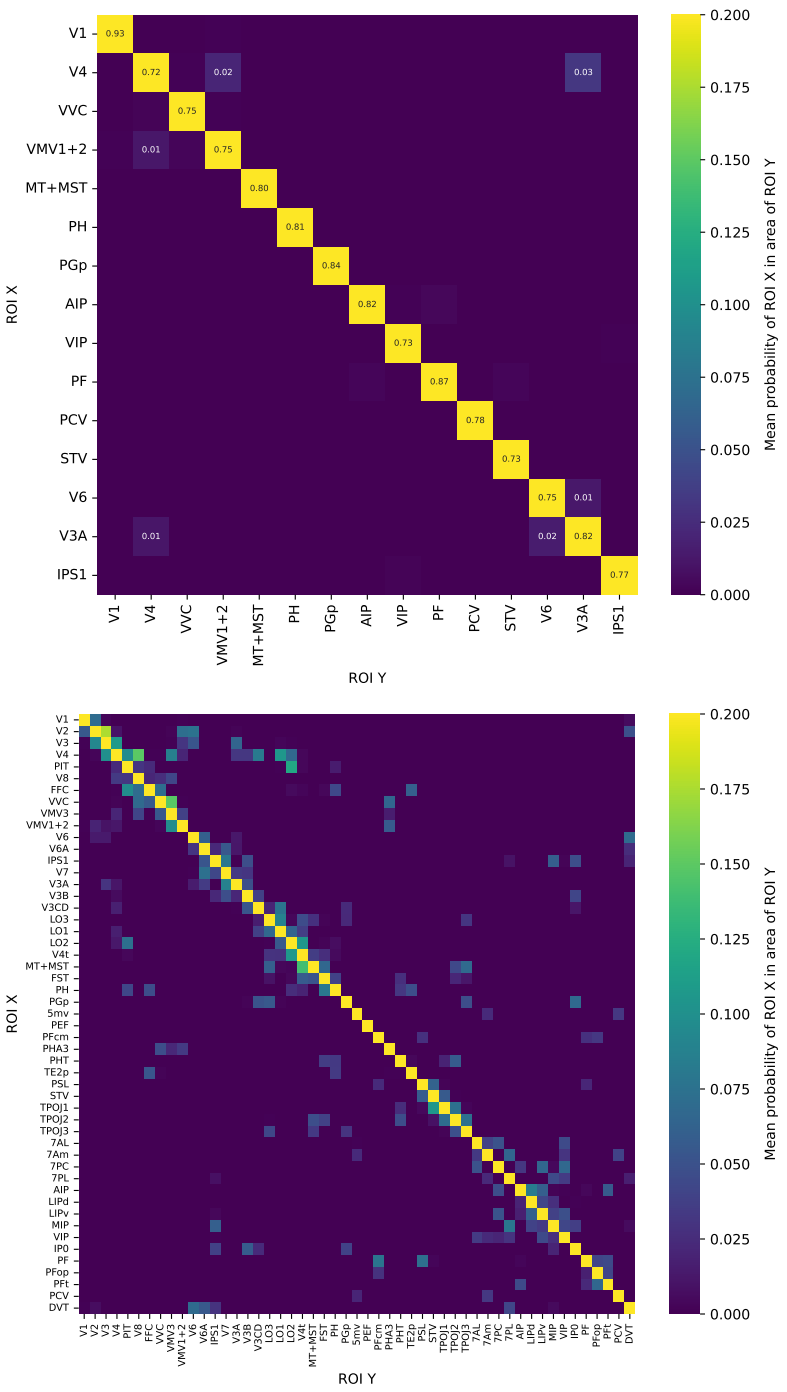


Figure 9: **Top:** Region overlap in the probabilistic map of the Glasser atlas (Glasser et al., 2016), for regions shown in figures 7. **Bottom:** Similar overlap matrix for all 'stimulus responsive' regions in BMD, for an overview of typical overlap probabilities between adjacent regions. Same color code.

shows the influence of de-meaning, i.e., subtracting the mean across stimuli, per voxel, on Pearson- and Euclidean RSA.

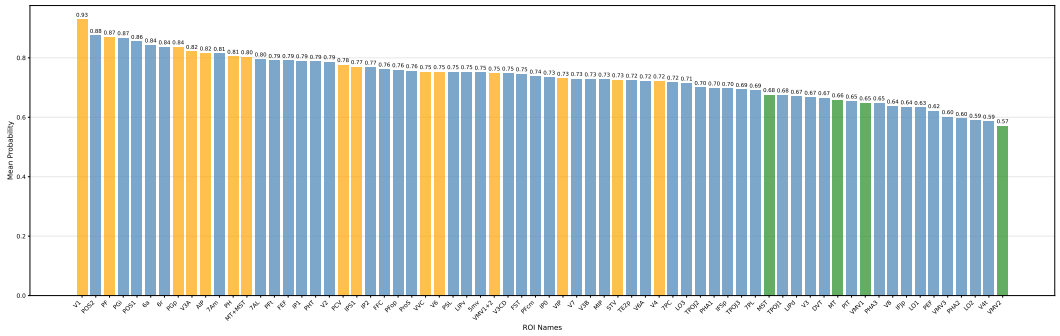


Figure 10: Mean probability of each ROI inside its region. Values computed from the probabilistic (validation) map of the Glasser atlas (Glasser et al., 2016). ROIs included

RSA: Spearman vs Kendall rank correlation

We always use Kendall's tau, a rank correlation method, to compare RDMs in all RSA variants. (We use SciPy's $scipy.stats.kendalltau$.) Figure 22 shows that for our purposes, Kendall's tau leads to the same results as using Spearman's rank-order correlation coefficient.

Measure agreement: Cohen's κ

Since it is unlikely that a DNN will ever perfectly match data from a brain region, we want to know whether measures agree on the similarity structure close to, but not at, the target. Therefore, we look at whether measures judge the same region as most similar to a target region, when the target region itself has been removed from the set of candidates. We expect this to be more informative than error consistency (Geirhos et al., 2020), since error consistency can only use cases where at least one measure makes an error.

To quantify agreement on the next-most similar region, we compute a variant of Cohen's κ. Cohen's κ (McHugh, 2012) is a metric of inter-rater agreement, correcting for the agreement expected by chance. Given observed agreement p_o (the fraction of cases where labels matched) and agreement as expected by chance p_e, Cohen's κ is then

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

i.e., how much more frequently two 'raters' agree compared to chance, set into relation to the gap between chance and the maximum possible agreement of 1. In our case, the measures are the 'raters' and the cases on which we judge the agreement are the target regions, to which the measures assign source regions as 'labels'.

The observed agreement is

$$p_o = \frac{1}{R} \sum_{r=1}^{R} \delta_{a_r = b_r},$$

where a_r is the region of maximum similarity to region r under the similarity measure A, b_r the region of maximum similarity to r under the other measure B, $a_r \neq r, b_r \neq r$, R is the number of target regions, and $\delta_{a_r=b_r}$ is the indicator function.

Since we exclude specific labels, different per case, we cannot use the standard computation of p_e which would be

$$p_{e,classical} = \sum_{r'=1}^{R} q_A(r')\, q_B(r') = \frac{1}{N} \sum_{i=1}^{N} \sum_{r'=1}^{R} q_A(r')\, q_B(r').$$

This takes into account how frequently each 'rater', or measure in our case, selects each class (or region) r': $q_A(r') = \frac{1}{N} \sum_{i=1}^{N} \delta_{a_i = r'}$.

Since we drop one label, and a different label per case (the target region is excluded as label), $q_A(r')$ as given above is not the correct marginal probability of choosing region r'. Instead, we compute

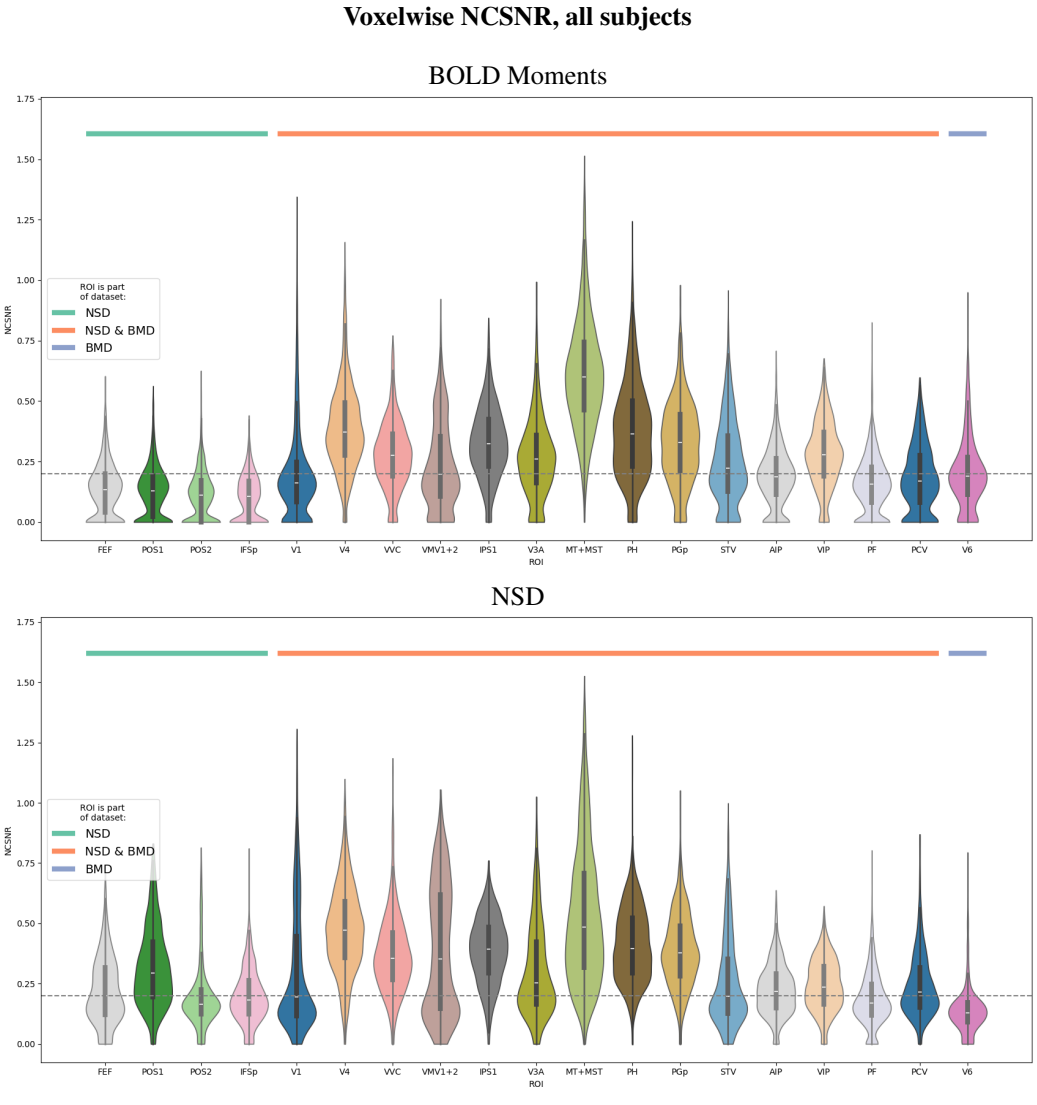


Figure 11: Voxelwise NCSNR distribution across ROIs (violin plots). Shown are distribution over all subjects jointly. Horizontal bars indicate for which dataset each ROI was included in analyses - V6 was only analyzed on BMD.

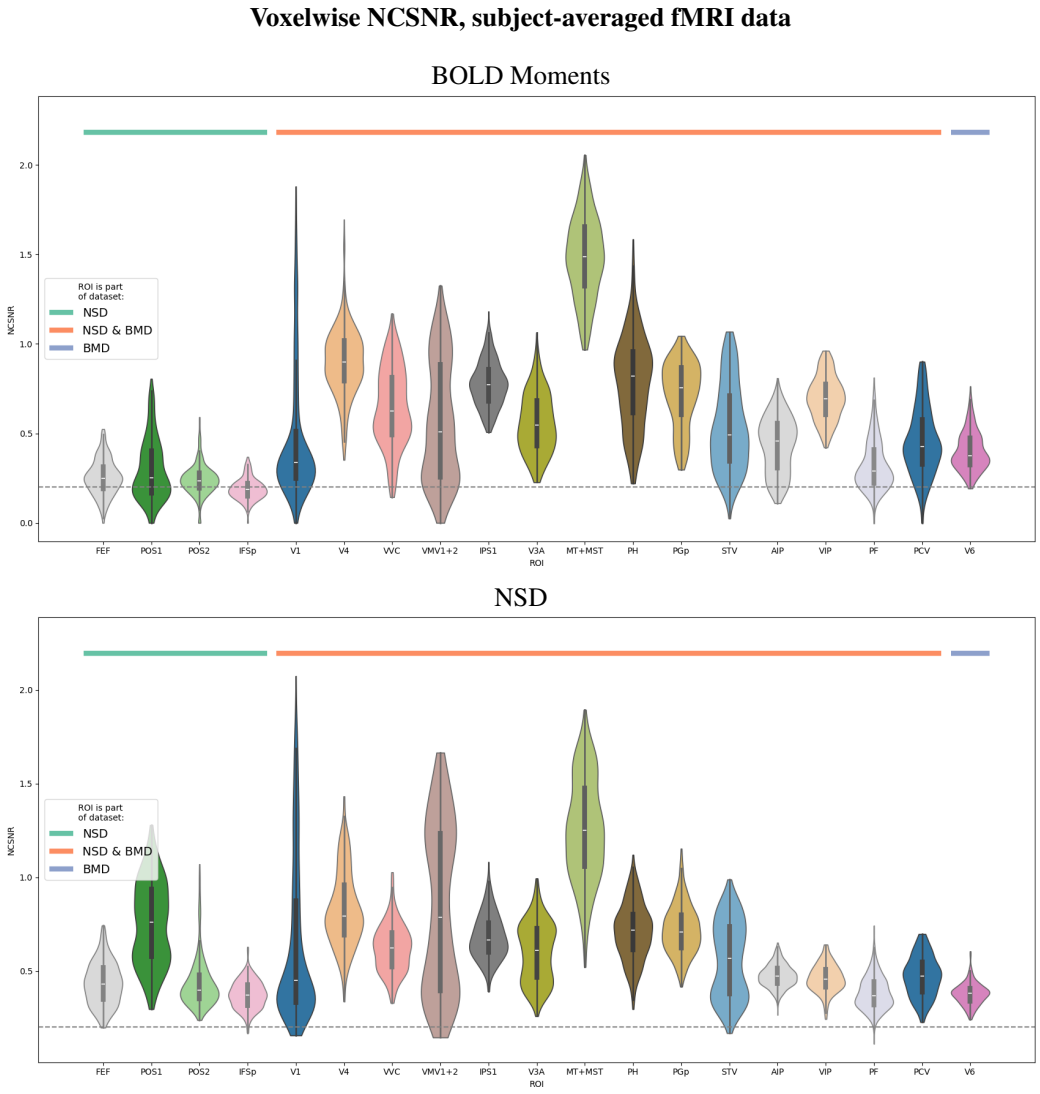


Figure 12: NCSNR distribution of the *subject-averaged* fMRI data (violin plots). That is, we average the fMRI beta value maps across subjects, for each stimulus, and then recompute the NCSNR of the averaged data, per voxel. Note the slight increase in y-axis range compared to fig. 11. Averaging improves NCSNR, showing the advantage of using leave-one-out averaged data as reference 'model'.

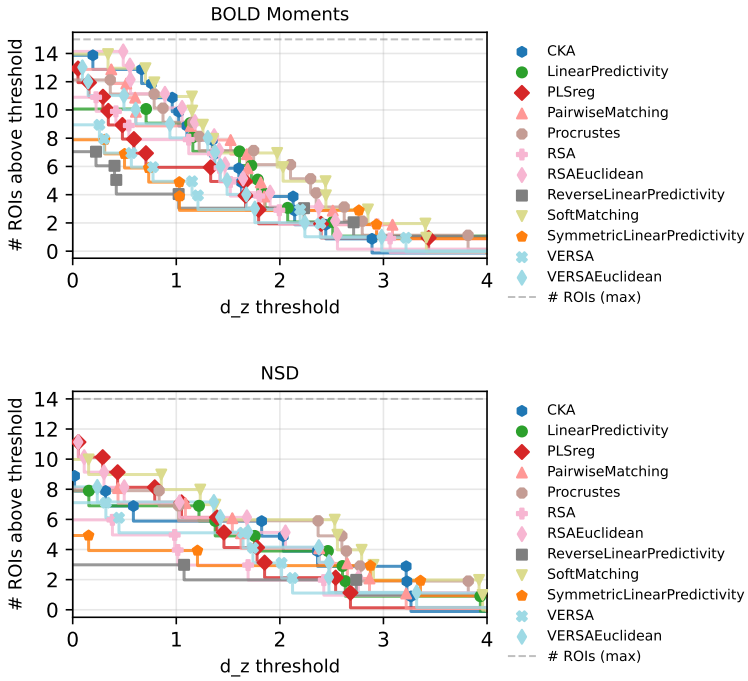


Figure 13: **Region identification based on varying thresholds of d_z** How many regions can a measure identify, if we consider a given d_z value on the x-axis as threshold? **Top:** BMD, **bottom:** NSD.

the marginal frequency with which a measure selects a region as most similar non-matching region, across all other regions. Since these marginal frequencies, computed naively, do not sum to 1, we need a target-region specific normalization term Z_A.

That is, we compute chance agreement probability as

$$p_e = \frac{1}{R}\sum_{r=1}^{R}\sum_{r'\neq r} p_{r,A}(r')\,p_{r,B}(r')$$

with $p_{r,A}(r') = \frac{1}{Z_A(r)}\tilde{p}_A(r')$, where $\tilde{p}_A(r') = \frac{1}{R-1}\sum_{i\neq r'}\delta_{a_i=r'}$. That is, $\tilde{p}_A(r')$ is the fraction of regions $\neq r'$ for which r' was the closest non-matching region. $Z_A(r)$ is a region-specific normalization factor, since the $\tilde{p}_A(r')$ do not sum to 1:

$$\sum_{r'\neq r}\tilde{p}_A(r') = \sum_{r'=1}^{R}\tilde{p}_A(r') - \tilde{p}_A(r) = \left(\frac{1}{R-1}\sum_{r'=1}^{R}\sum_{i\neq r'}\delta_{a_i=r'}\right) - \tilde{p}_A(r) = \frac{R}{R-1} - \tilde{p}_A(r)$$

by exchanging summation order. That is, $Z_A(r) = \frac{R}{R-1} - \tilde{p}_A(r)$.

Similarly $p_{r,B}(r') = \frac{1}{Z_B(r)}\frac{1}{R-1}\sum_{i\neq r'}\delta_{b_i=r'}$.

Finally, given p_o and p_e, Cohen's κ is then, as in the classical case, $\kappa = \frac{p_o - p_e}{1-p_e}$.

Disentangling univariate contribution to each measure

Measures applied to only the scalar region-average (univariate) signal perform close to the top performing measure in the standard setting (fig. 28 left). When the univariate signal is removed (fig. 28 center), all measures' region matching performance is decreased.

This raises questions: If measure performance is heavily influenced by the univariate signal component, are they suitable for evaluating similarity of DNN models to brain areas?

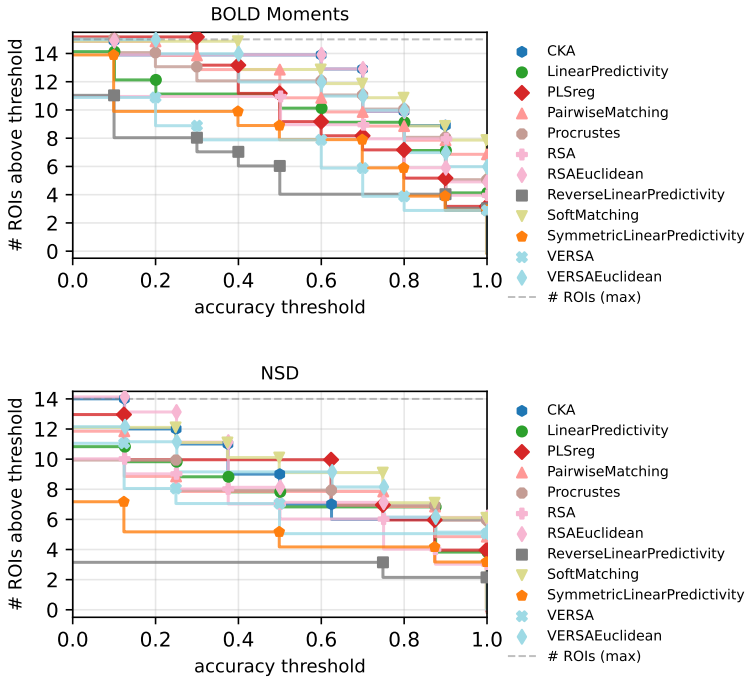


Figure 14: **Region identification based on varying thresholds of region-matching accuracy. Top:** BMD, **bottom:** NSD.

Univariate activity in DNNs is unlikely to be comparable to univariate activity in the brain: Univariate activity in DNNs is not generally assumed to carry information. On the opposite: A commonly used component, Layer Normalization (Ba et al., 2016) removes the preceding layers' mean activity, meaning their mean cannot carry functionally relevant information. Further, transformer models have been found to have massive-magnitude 'rogue dimensions' that are mostly constant (Timkey & Van Schijndel, 2021; Sun et al., 2024), but have a functional role, which will dominate the mean. Their influence can be countered by standardization at the time of similarity computation (Timkey & Van Schijndel, 2021) However, they cause the networks mean at inference time to be close to constant. Therefore this mean carries a different functional role in DNNs compared to fMRI data.

This highlights a potential issue in brain-DNN similarity benchmarks: Similarity scores can be driven by signal components that provably carry divergent function in brains compared to DNNs. It is not generally the case, however, that a subcomponent of the signal has to match between source and target, for a measure to use it to evaluate similarity - Euclidean RSA can match similarity of patterns to similarity of a scalar signal, and LP can predict a scalar signal from a source that has its univariate signal removed. For other measures, closing the 'modality' gap before similarity scoring might help to a certain degree (Kriegeskorte & Diedrichsen, 2019). fMRI carries a fundamentally coarse signal, commonly thought to represent the blood flow to a part of gray matter, corresponding to its activity level - but see (Epp et al., 2025). A voxel in fMRI, or a vertex in the processed surface map, is very unlike a computational unit in a DNN. E.g., the fsLR32k surface map (Van Essen et al., 2012) splits ca 16 billion cerebral cortex neurons (Azevedo et al., 2009) into ca 60'000 vertices, so that a single vertex on average corresponds to roughly 270'000 neurons per vertex. It is questionable that the pattern of such coarse activity will be 'similar' to those of low-level computational units in a DNN, even if the latter were to match the brain's computation.

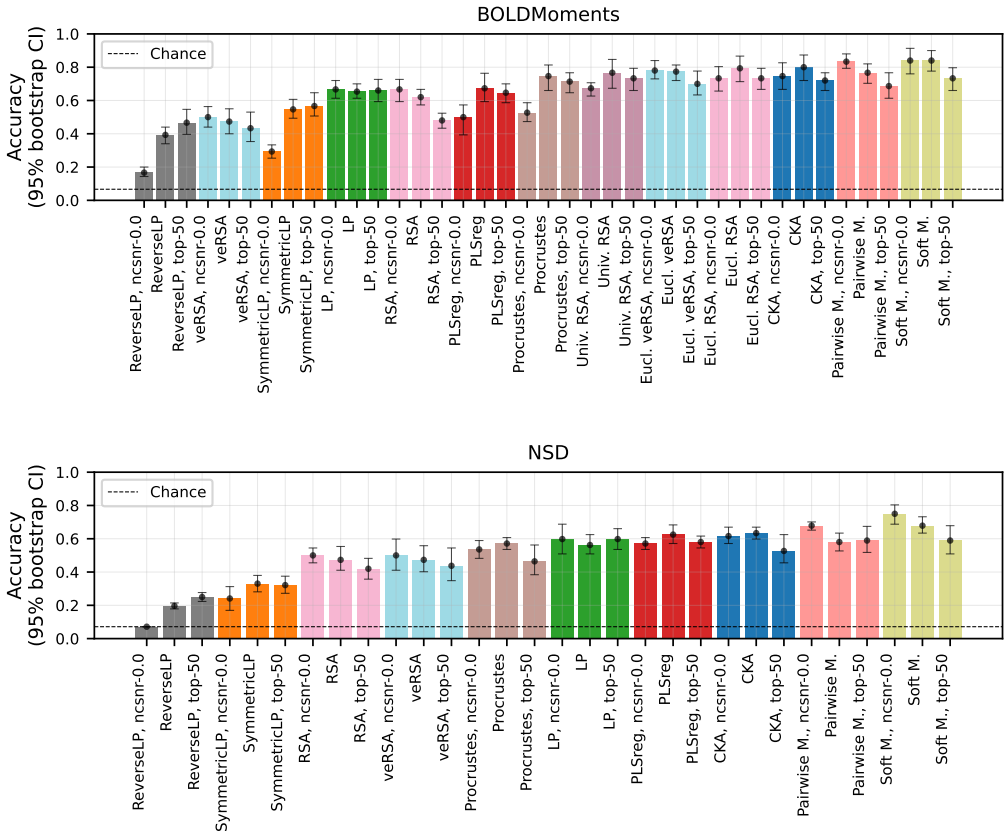


Figure 15: Comparison of three voxel selection methods across a range of measures, **top:** on BMD, **bottom:** on NSD. Noise threshold filtering is helpful or has no effect for most higher-performing measures.

Feature specificity

The unreasonable effectiveness of SoftMatching on regions that are hard to identify

Shown in fig. 33

Measures fundamentally disagree, even on regions they can identify

Measures have been found to disagree strongly (Soni et al., 2024; Huh et al., 2024). If measure agreement is evaluated on regions which they cannot distinguish well, noise or spurious results might inflate disagreement. However, on two measures of agreement (section H.3.1), dropping regions below an identifiability threshold increases agreement only slightly or not significantly (figs. 23 and 24).

Generally, measures disagree considerably, even without any brain-DNN modality gap present (fig. 26). Chance-corrected agreement κ is around 0.4, i.e., on average in 40% of remaining cases that would disagree in a chance setting, two measures agree. That is, not even on brain data, measures agree on the similarity structure close to the target region.

Measure agreement on identifiable regions: Method details

We compute agreement on the closest non-matching region as Cohen's κ, section H.1.5, and as Spearman rank correlation of similarity scores. The self-match is always excluded.

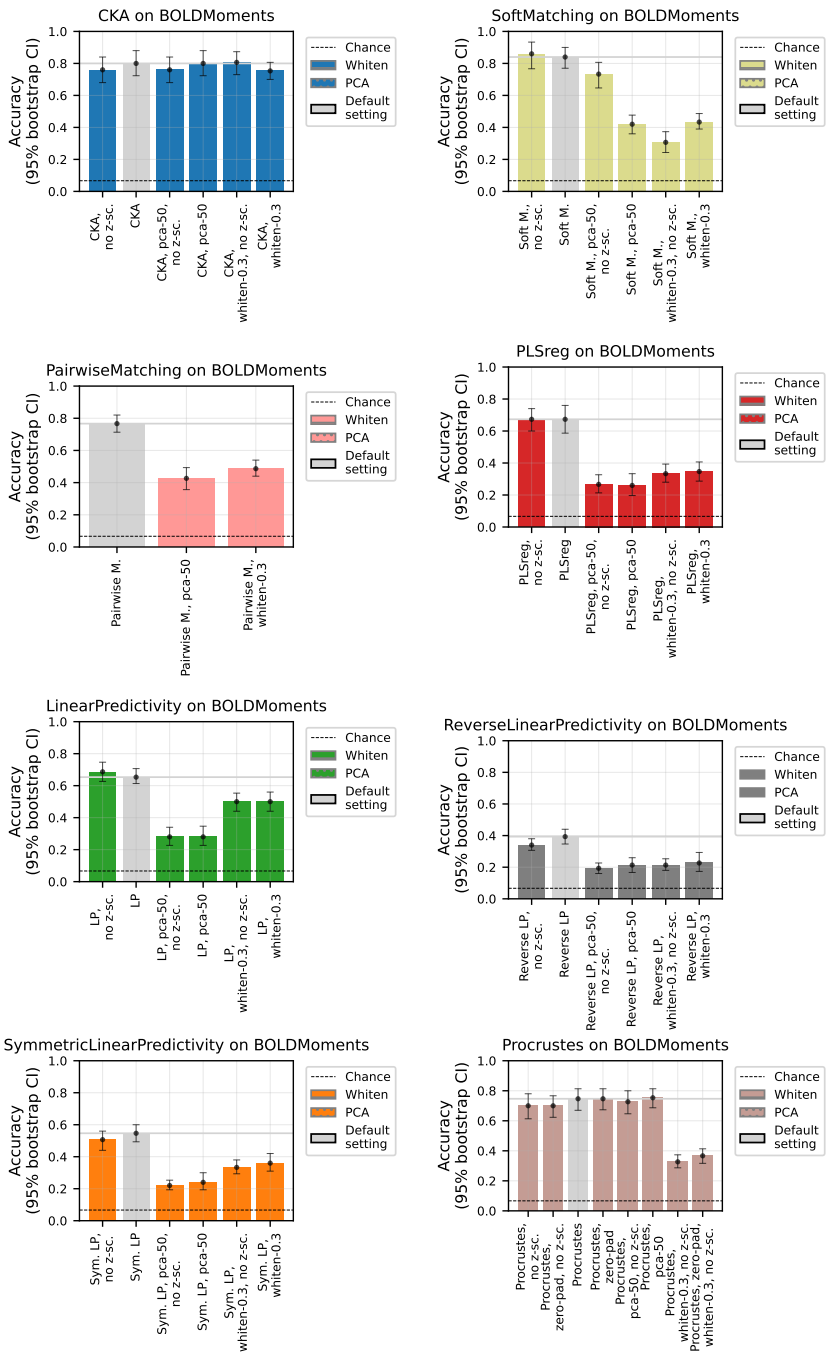


Figure 16: Region matching accuracy on BMD, for various measures, with different preprocessing methods. Z-scoring has little effect overall, while PCA and whitening have a detrimental effect on different sets of measures. We do not vary z-scoring for PairwiseMatching, since it is based on voxelwise correlations which do not change under z-scoring.

Measures included: We include Euclidean and Pearson-distance based RSA and veRSA variants, but no univariate variants, and two settings per measure, one with feature reduction to the first 50 principal components, one without. We include PCA-preprocessing variants for comparability with Soni et al. (2024); results without PCA preprocessing are showin in fig. 25. Heatmaps fig. 26 include only settings without PCA. Significance in threshold vs. agreement curves (figs. 23 to 25) is computed by comparing to dropping random regions, matched in number, to account for increases in chance

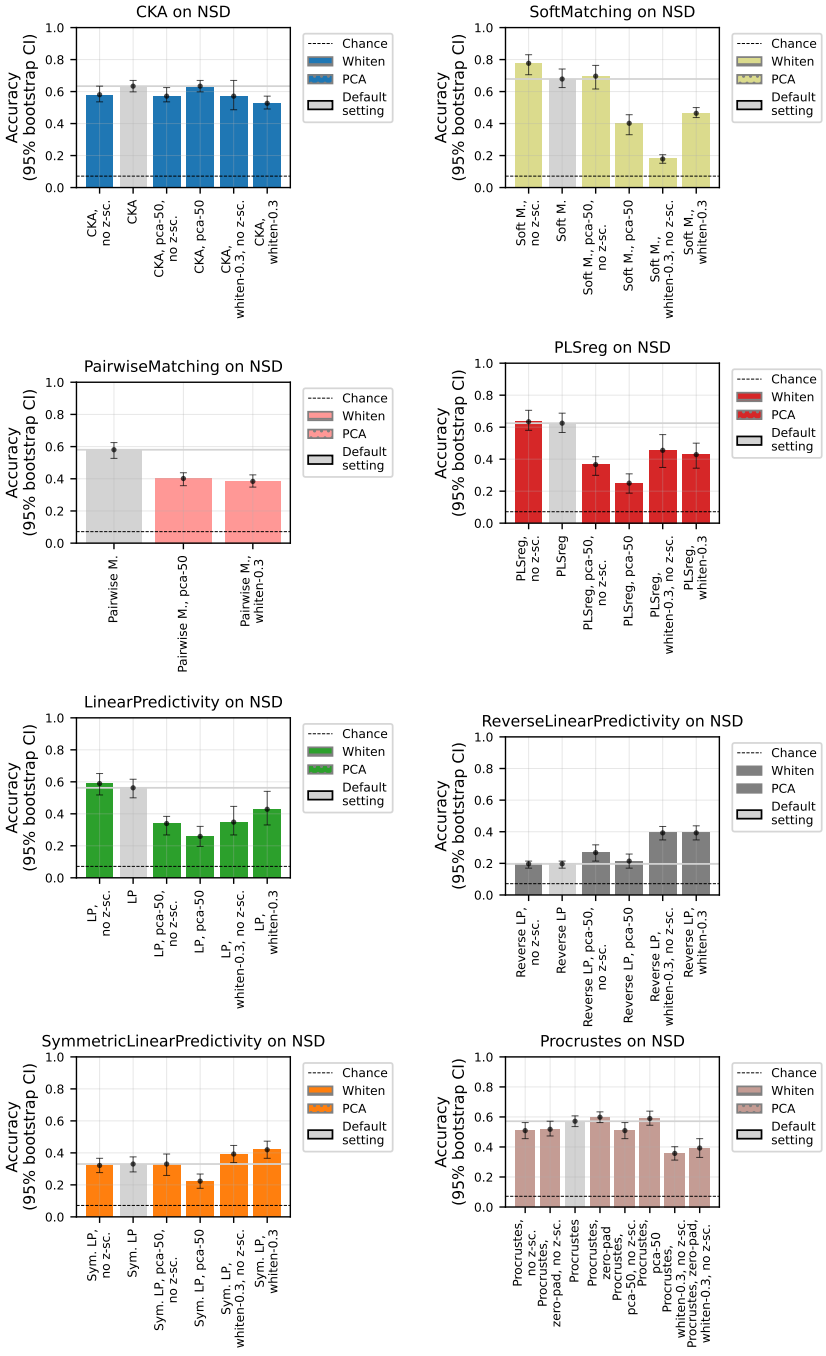


Figure 17: Region matching accuracy on NSD, for various measures, with different preprocessing methods.

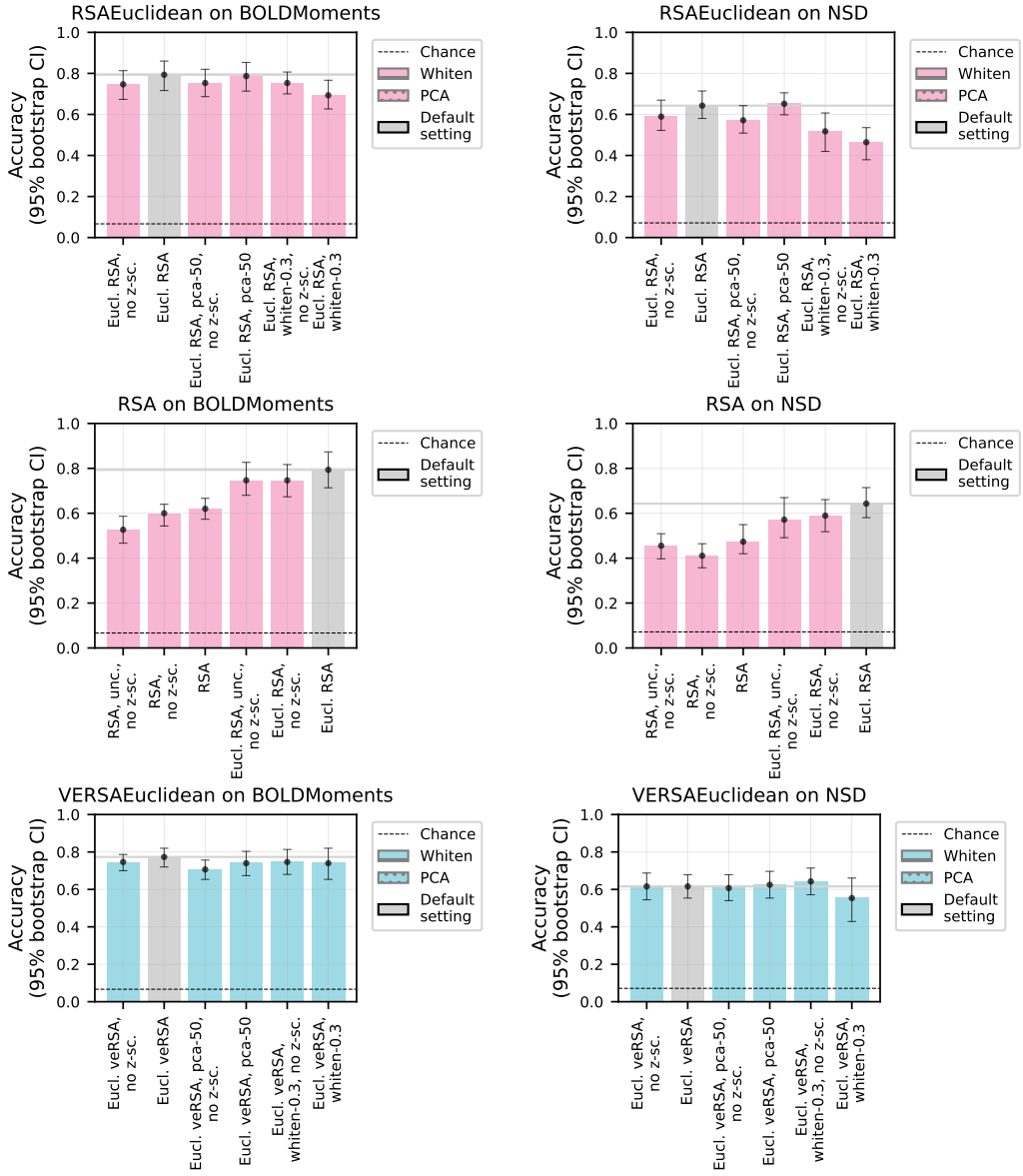


Figure 18: Region matching accuracy on BMD, for Euclidean (ve)RSA methods, with different preprocessing methods. Center row ablates de-meaning (Ramírez, 2017) - no de-meaning: 'unc.' for 'uncentered' - for both P-RSA and Euclidean RSA.

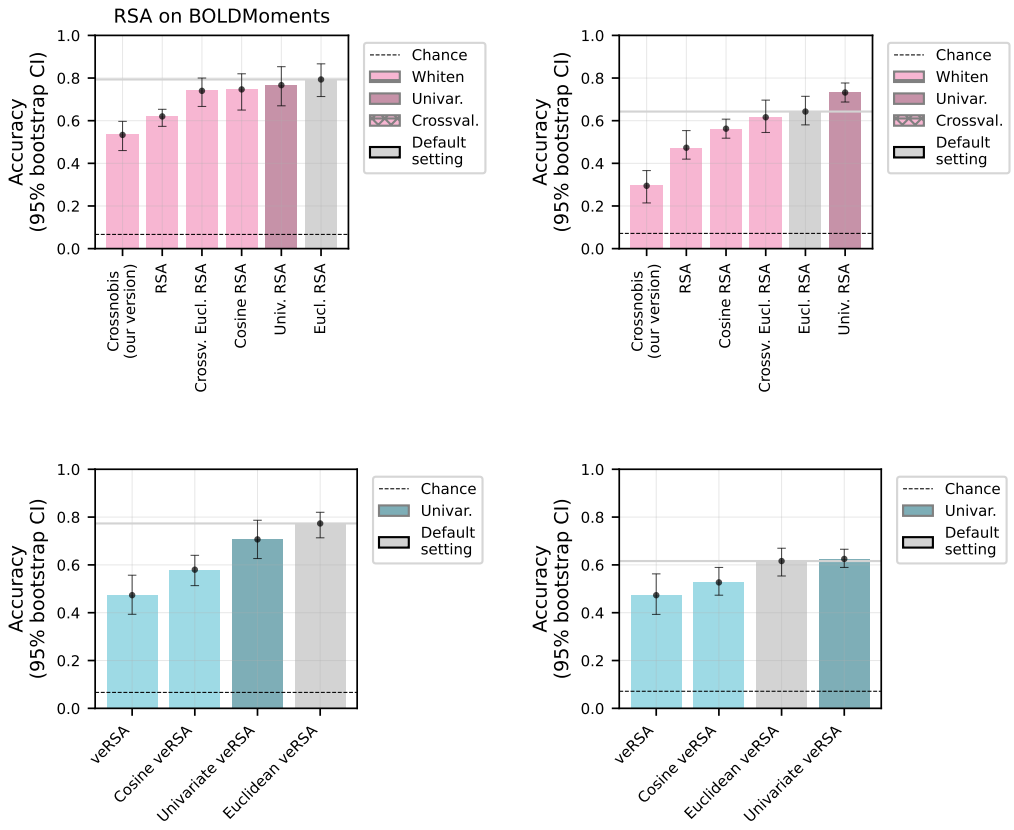


Figure 19: **ROI matching accuracy: RSA and veRSA distance variants Left:** BMD, **right:** NSD. **Top:** RSA, **bottom:** veRSA. Top left figure is the same as fig. 4

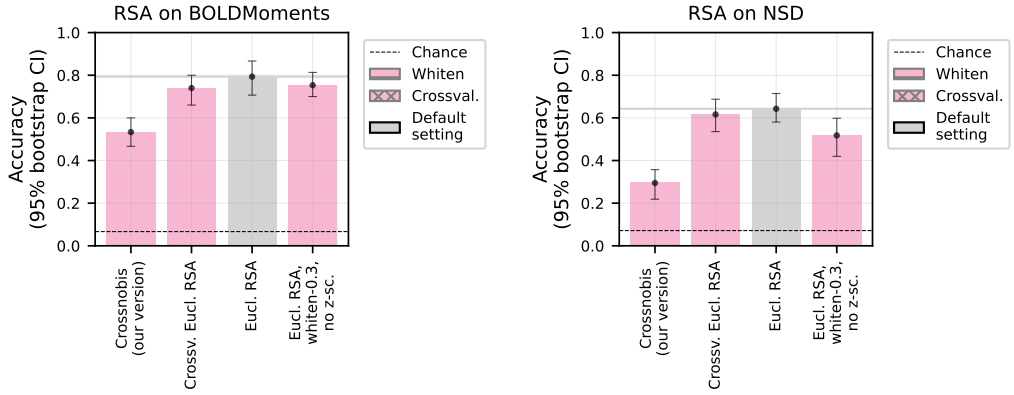


Figure 20: **Crossvalidation and whitening disentangled**. ROI matching accuracy: Euclidean RSA variants with and without crossvalidation. **Left:** BMD, **right:** NSD. It seems to be mainly the combination of crossvalidation with whitening ("Crossnobis (our version)") what reduces performance.

agreement due to fewer available regions. In these plots, shaded areas are 95% bootstrap confidence intervals, the shaded region around the main curve is using subject resampling, the nulls' CI region is computed from both subject and randomly-dropped region resampling.

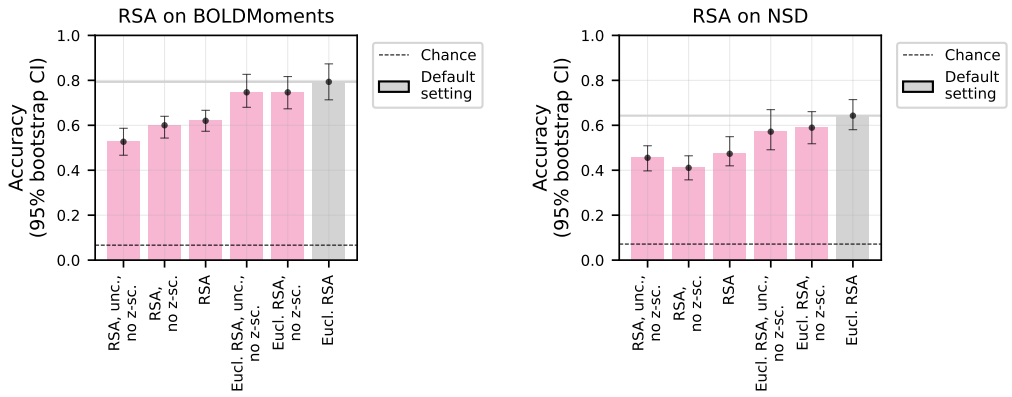


Figure 21: **Centering, in Pearson and Euclidean RSA**. **Left:** BMD, **right:** NSD. **'unc.'**: uncentered, **no z-sc.**: no z-scoring, i.e. no normalization across stimuli per voxel. Centering or not seems to have no noticeable impact on region matching performance, while RSA (i.e., the Pearson-distance variant) underperforms Euclidean-distance RSA.

VALIDATION EXPERIMENTS

MATCHING LOW-NCSNR REGIONS

To test whether the measures we test are picking up on attributes of the data that are *unrelated* to the stimulus-induced signal, we test the measures on region-matching of regions with low NCSNR. For this we first select the set of 50 regions with the highest fraction of voxels below 0.1 NCSNR. Then we manually exclude regions TE2a, FOP1, TE1m, EC, H, AVI, and MI, since they appeared to potentially have a visual function. We missed to exclude PreS (Pre-Subbiculum). We force-include OFC and region 4, to ensure a few large regions are contained in the sets. We use integer optimization to select a subset, maximizing the number of regions included while keeping region overlap, based on Glasser-atlas region probability maps, at or below 0.3. We use the CBC (Coin-or branch-and-cut) solver implementation from the $PuLP$ (Python LP Modeler) Python package (class PULP_CBC_CMD, (Forrest et al., 2024)), to determine two distinct sets of "noisy regions" for BMD and NSD. Tables 4 and 5 show the selected regions with their number of voxels, fig. 34 shows their Glasser-probability overlap, and fig. 35 their distributions of NCSNR values.

ROI	Num. voxels
PreS	328
OFC	680
Pir	383
10d	271
8BM	349
SFL	304
47s	265
p47r	229
8Ad	481
4	1649
STSva	210
24dv	298
10r	168
p9-46v	309
3b	1131

Table 4: BMD low-NCSNR regions: Number of voxels per ROI

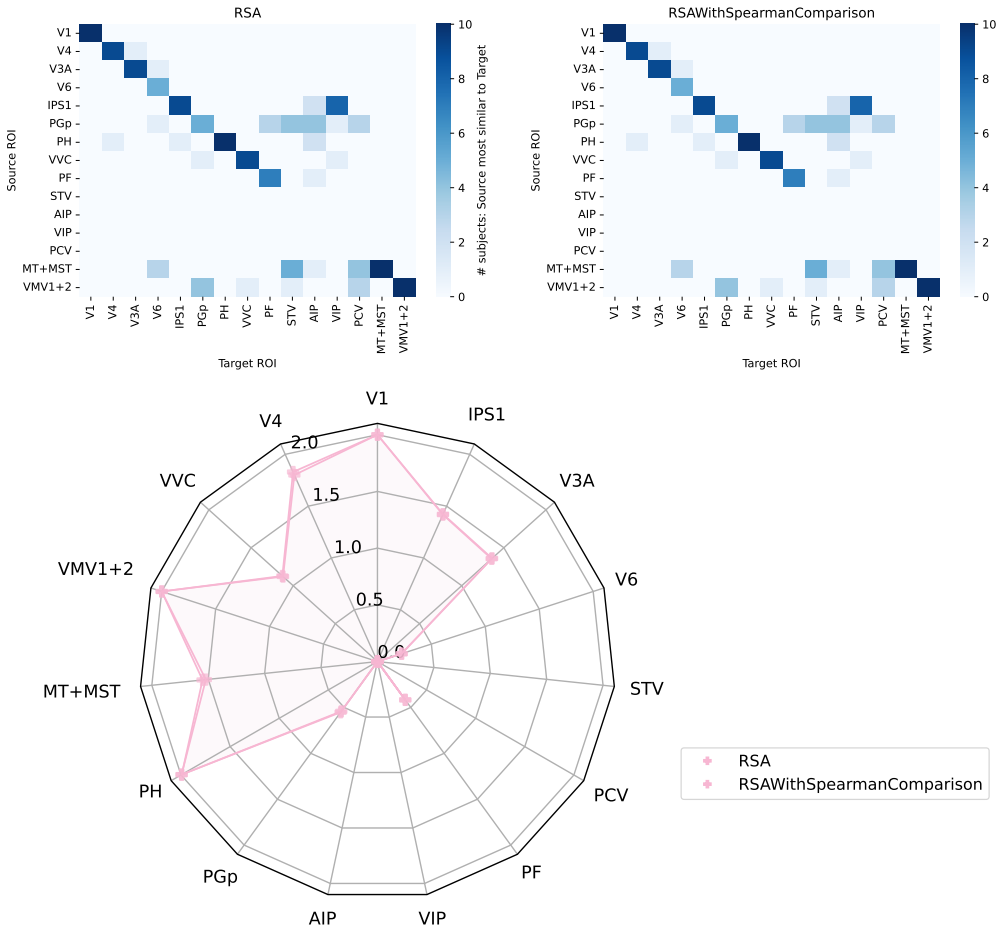


Figure 22: **RSA, Kendall vs. Spearman correlation Top:** Confusion matrices for RSA using Kendall tau to compare RDMs (this is our default RSA implementation; right), and for RSA with Spearman r (left). **Bottom:** d_z values across ROIs, for the two RSA variants. Even here, differences are tiny, curves overlap almost completely.

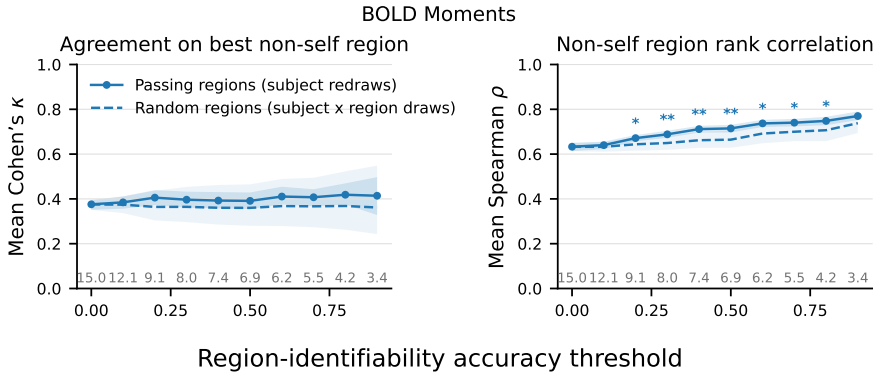


Figure 23: **Measure agreement rises slightly on identifiable regions (BMD).** Agreement between pairs of similarity measures, computed only on regions where both measures exceed the identifiability threshold on the x-axis (passing regions), compared to a null of dropping the same number of random regions. Details in section H.3.1. **Left:** agreement on the closest non-matching region. **Right:** Spearman rank correlation. Gray numbers show average number of regions retained across measure pairs.

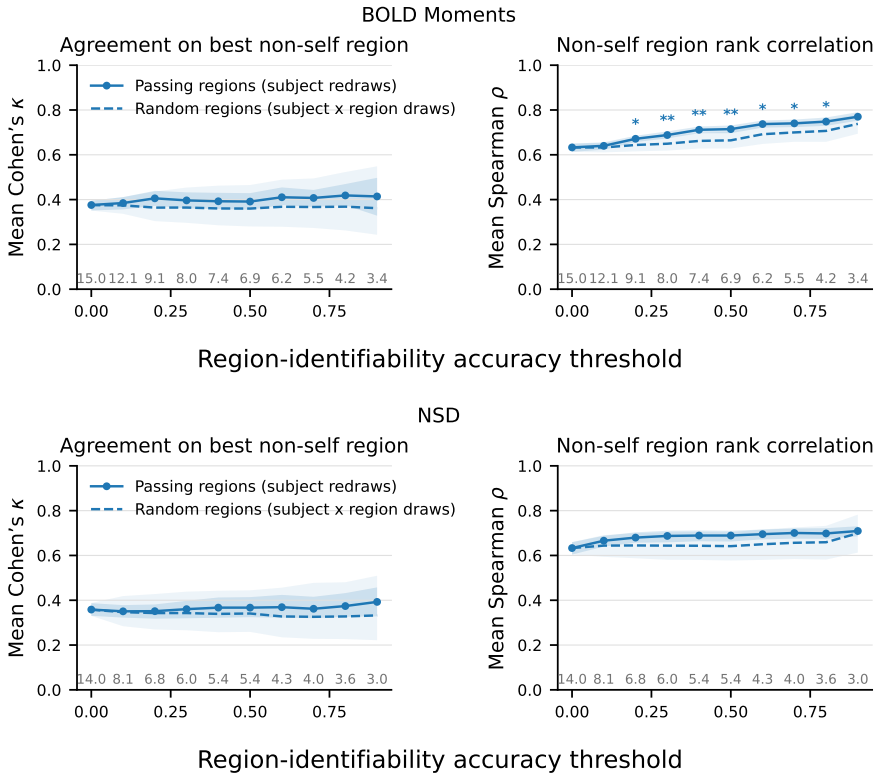


Figure 24: **Measure agreement rises slightly on identifiable regions. Top:** BMD (repeated figure **??**) **Bottom:** NSD. Each plot based on 24 measure configurations, 12 with feature reduction to the first 50 principal components, 12 without feature reduction. Agreement between pairs of similarity measures, computed only on regions where both measures exceed the identifiability threshold on the x-axis. **Left:** agreement on the closest non-matching region (Cohen's κ). **Right:** Spearman rank correlation of similarity scores. The self-match is always excluded. **Measures included:** "We include Euclidean and Pearson-distance based RSA and veRSA variants. Two variants per measure, one with feature reduction to the first 50 principal components, one without. Gray numbers show average number of regions retained. Significance is computed by comparing to dropping random regions, matched in number. Shaded and dashed area are 95% bootstrap confidence intervals, the shaded region using subject resampling, the dashed (null) region using both subject and randomly-dropped region resampling.

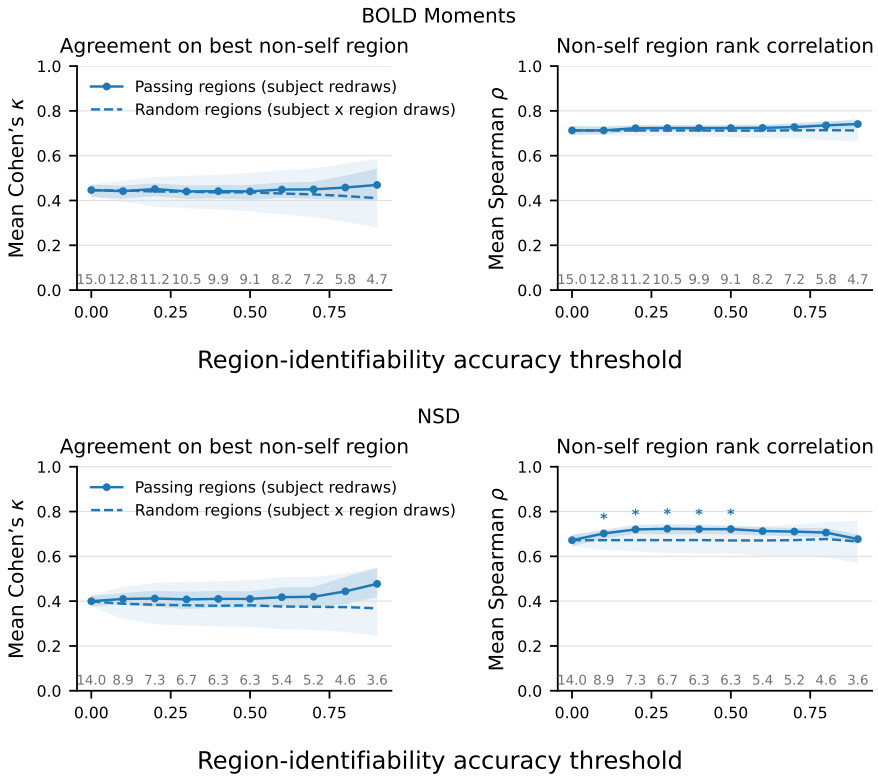


Figure 25: As figure 24, but averaging only across all pairs of the 12 measure configurations without PCA-based feature reduction. **Top:** BMD, **bottom:** NSD. The increase in agreement is much smaller, indicating that PCA-preprocessed settings carry much of the unidentifiability-related measure disagreement.

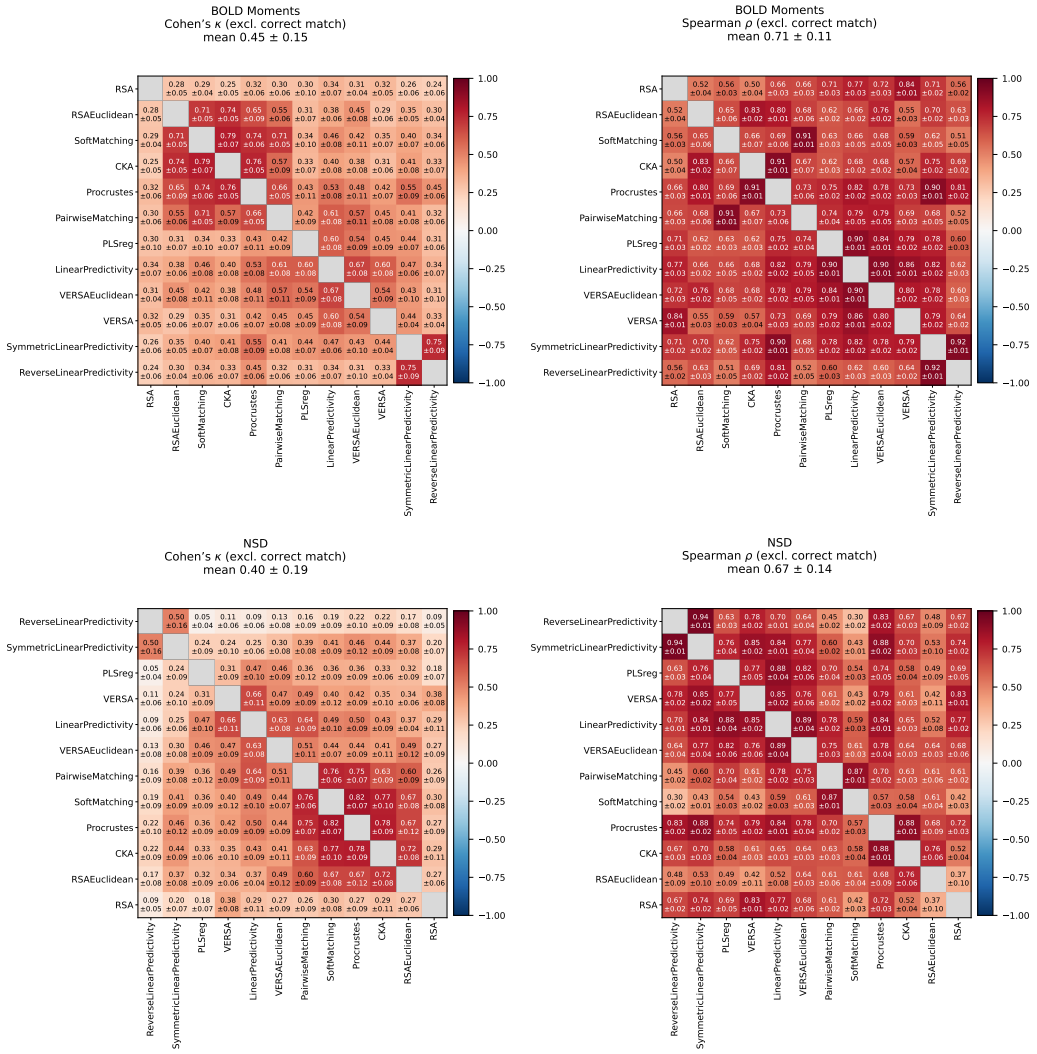


Figure 26: **Agreement between measures:** Pairwise agreement values entering fig. 25 (no PCA). Values after +− are half the width of the 95% confidence interval, from resampling subjects. +− in the title shows standard deviation of agreement values across measure pairs.

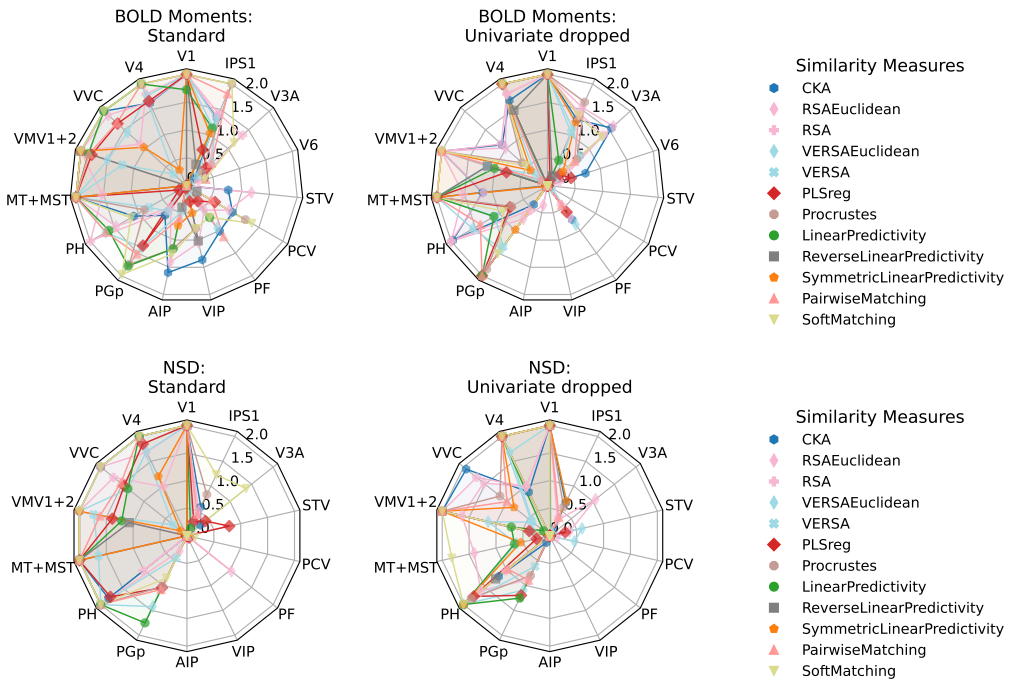


Figure 27: Removing each region's average response per stimulus lowers region identifiability for most similarity measures. For BMD, many associative regions drop to undetectability when the univariate signal is removed. Shown: minimum d_z over all distractor source regions, clipped to [0, 2]. Left: standard setting, for comparison. Right: the region's mean response per stimulus is removed from both feature maps before the measure is applied. **Top row:** Bold Moments; **bottom row:** NSD.

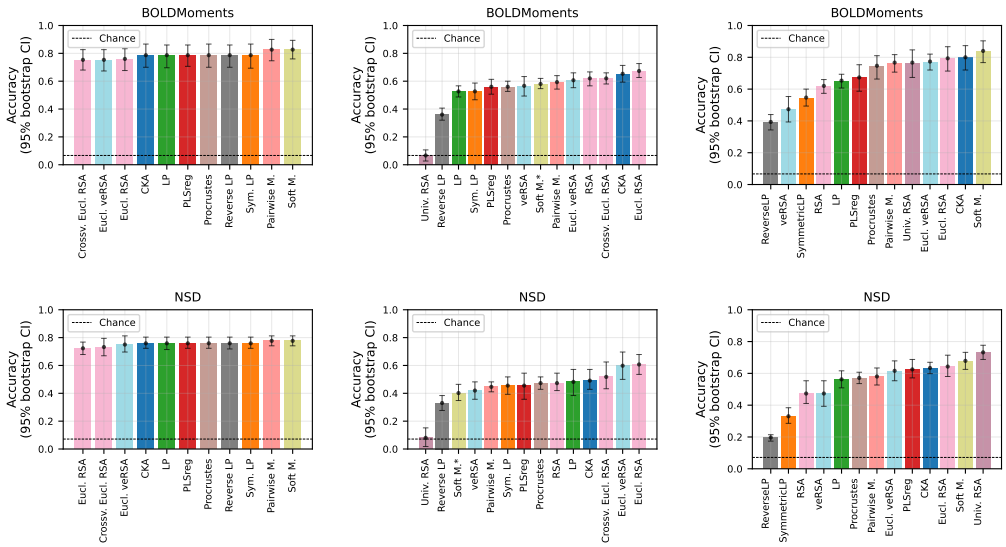


Figure 28: **Only univariate information, no univariate information. Left:** Measure performance when applied to the mean voxel beta value per stimulus, instead of per-stimulus beta (activity) *vectors*. **Center:** The reverse - measure region matching performance on data where mean activity per stimulus has been subtracted. **Right:** Original measure performance, for comparison. Note some measures collapse to the same measure under univariate-only application - such as SoftMatching andPairwiseMatching, and PLSreg (reducing to Linear Regression). It is striking how all univariate measures perform close to the top performing measure overall. **Top:** BMD, **Bottom:** NSD. Note: There seem to be correlated floating point inaccuracies, where we find RSAUnivariate scores highly above chance when dropping the univariate signal, which is technically impossible - RSAUnivariate should be comparing all-zero RDMs. We therefore add uniform noise above the level of floating point accuracy after subtracting the per-stimulus mean. We use $\epsilon = 10\times$ float32 resolution (ca 1.12×10^{-6}) for NSD, and $\epsilon = 10^{-10}$ for BMD, and sample uniform noise from $[-\epsilon, \epsilon]$, ensuring that the absolute values of the data are much above this level of noise.

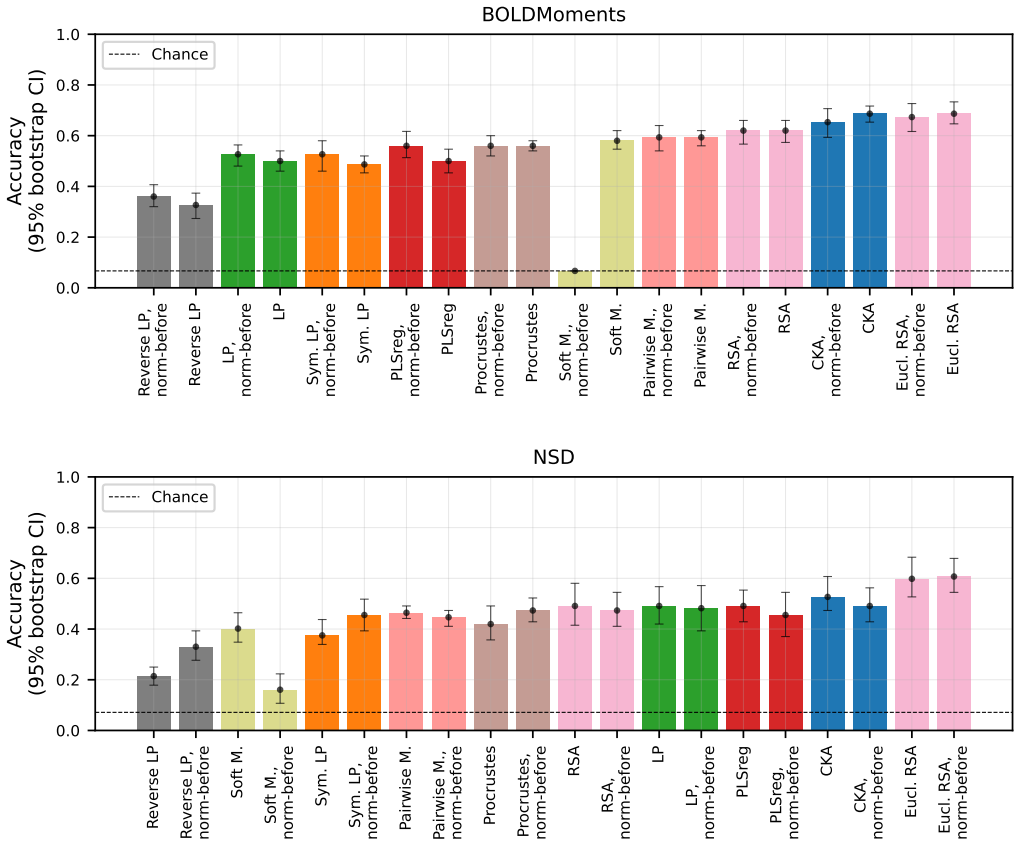


Figure 29: **Dropping univariate information either before, or after re-normalizing voxel activity** See also text, section section H.1.6. **Top:** Comparing region matching accuracy when renormalizing before ('norm-before') or after subtracting the univariate signal from the data, on Bold Moments. **Bottom:** Same for NSD. Most measures show only small differences, but accuracy using SoftMatching is severely impacted.

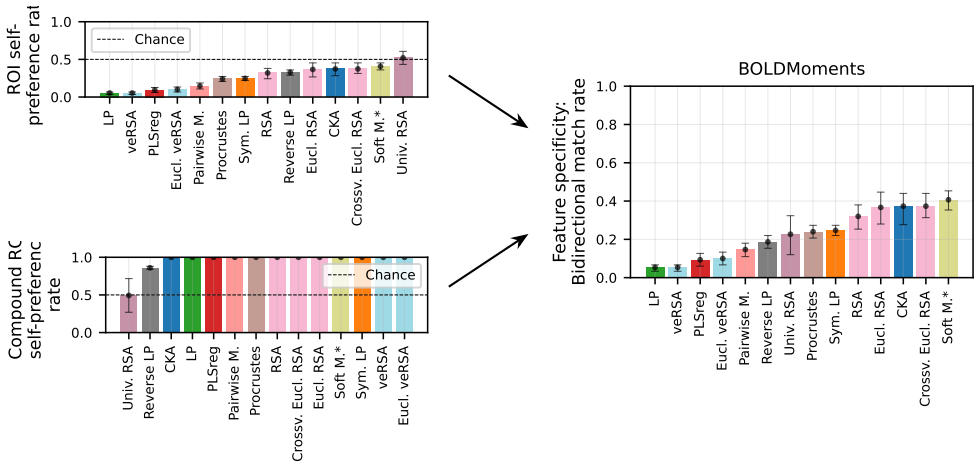


Figure 30: **Distinguishing regions from a compound region that contains them, without univariate information. BMD.** See also section 3.2.

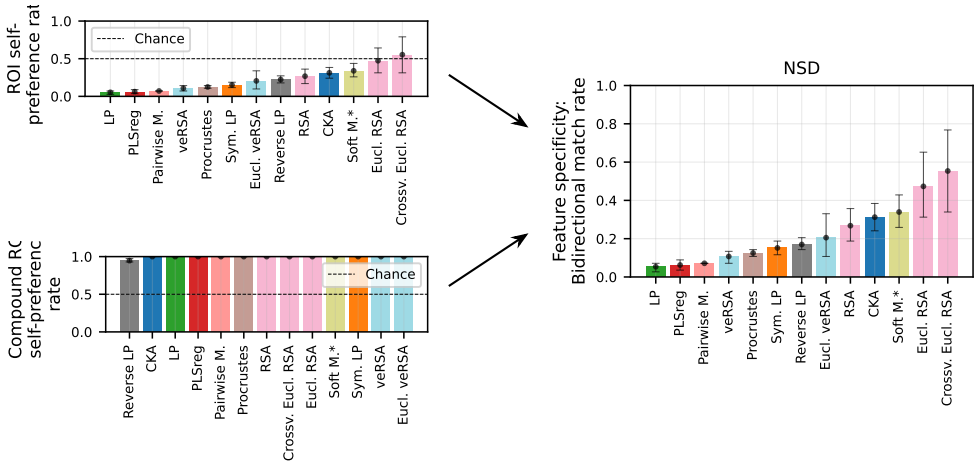


Figure 31: **Distinguishing regions from a compound region that contains them, without univariate information. NSD.** See also section 3.2.

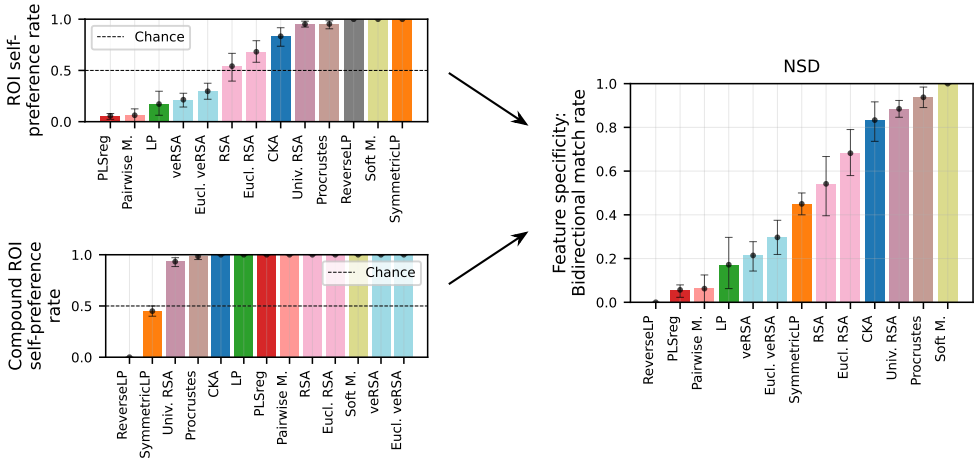


Figure 32: **Distinguishing regions from a compound region that contains them.** See also section 3.2. Same as figure 5 but for NSD.

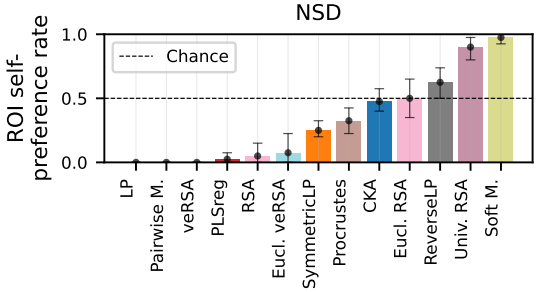


Figure 33: Fraction of subjects and regions where same-ROI similarity was larger than the similarity to the compound ROI, **for regions PF, STV, AIP, VIP and PCV on NSD.** Same type of figure as fig. 5 and section H.2, top left.

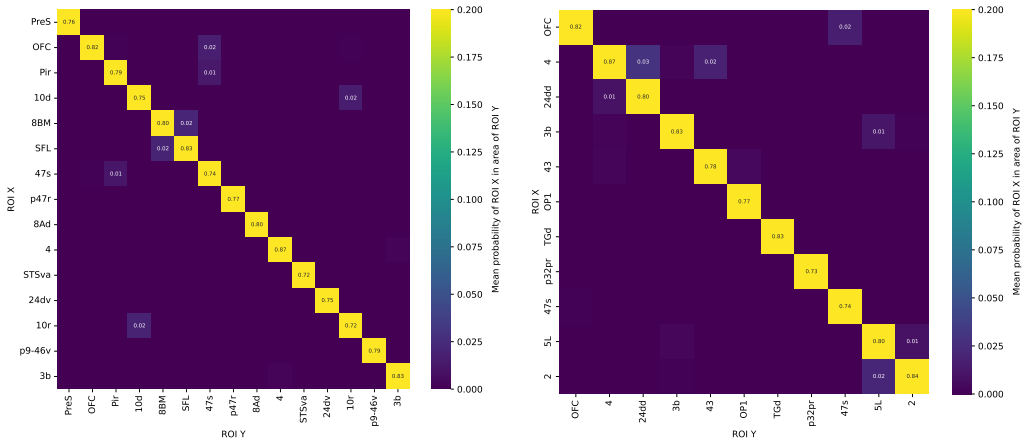


Figure 34: Glasser-atlas probability based anatomical overlap matrix, for noisy regions for BMD (left) and NSD (right).

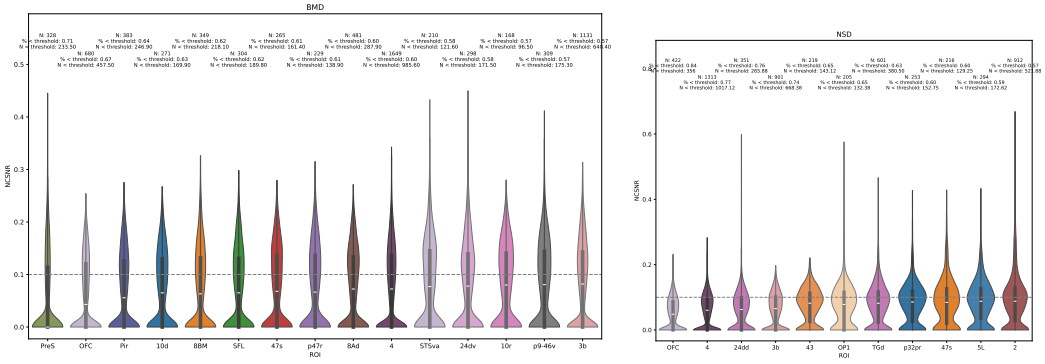


Figure 35: NCSNR distribution (all subjects jointly) for the noisy set of regions, for BMD (left) and NSD (right). Also annotated are the average number of voxels below an NCSNR threshold of 0.1. (Note this is a different threshold than when selecting "good" voxels, and we do not actually filter to only low-NCSNR voxels, this serves as a criterion to select noisy regions.

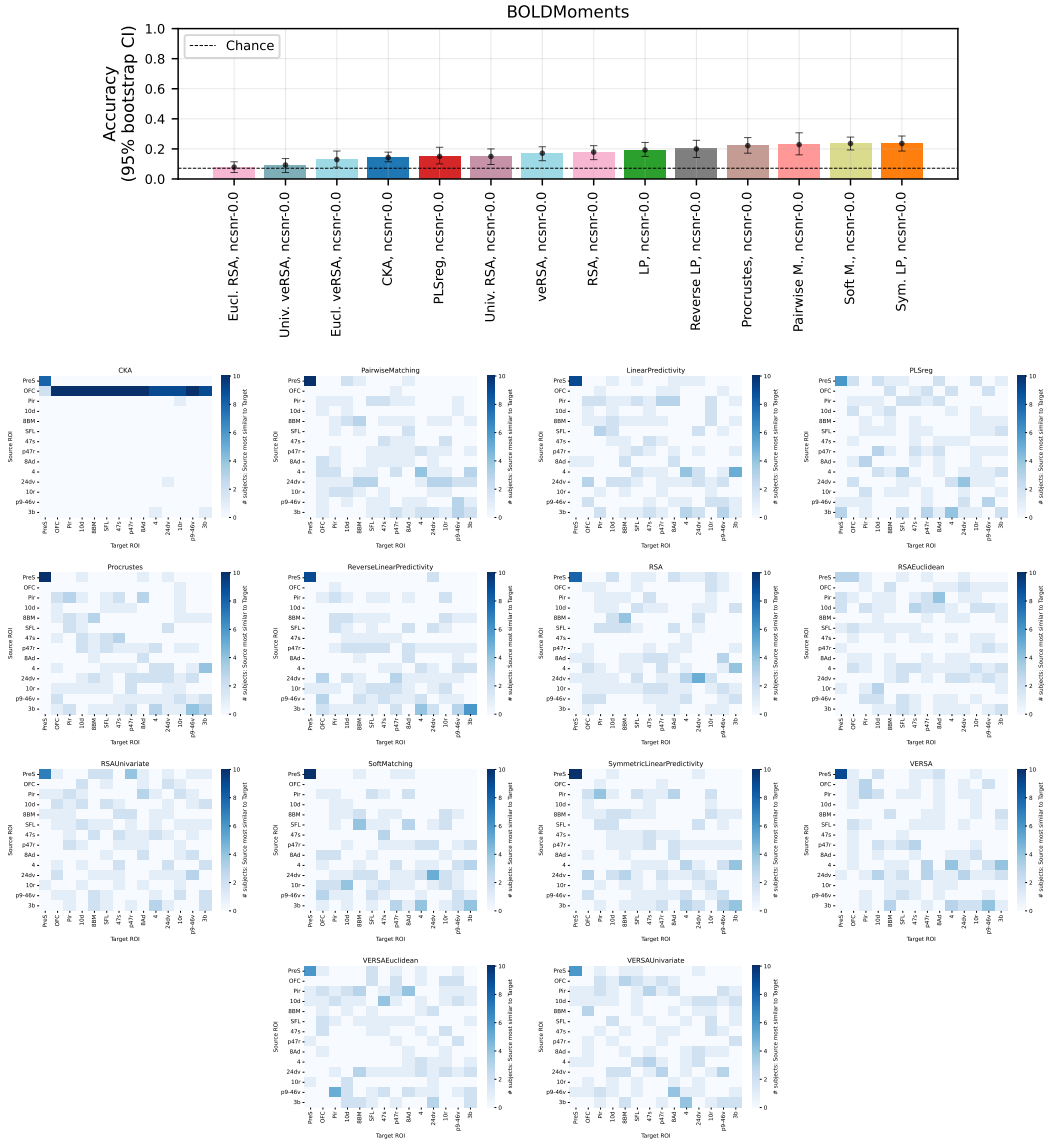


Figure 36: **Matching low-signal regions. Top:** Region matching accuracy on the set of noisy regions described in section I.1. **Lower part:** Region confusion matrices for all measures shown in the top figure. While individual measures appear to be able to detect a few regions above chance – beyond the clearly identifyable region PreS, which is a visual region that slipped our exclusion – if they could make use of confounders like ROI size, we would expect stronger patterns and better identifiability of, for example, regions with clearly distinct size such as OFC with ca. 700 voxels, region 4 with ca 1650 voxels, or region 3b with ca 1100 voxels. Although this experiment does not rule out confounders in the main regions we looked at, at least in the low-NCSNR (mostly) nonvisual regions shown here, there do not seem to be factors that help the measures identify a region, which are signal-unrelated. A striking pattern is, of course, CKA favoring the orbitofrontal cortex OFC, for all regions except PreS. This clear bias can at most hurt its performance, and in the presence of a signal (PreS), the bias seems to be overruled by other factors.

ROI	Num. vertices
OFC	422
4	1313
24dd	351
3b	901
43	219
OP1	205
TGd	601
p32pr	253
47s	216
5L	294
2	912

Table 5: NSD low-NCSNR regions: Number of vertices per ROI

RESULTS (NOISE-REGION MATCHING)

REGION-MATCHING UNDER EQUAL FEATURE MAP SIZE

In fig. 37 we show figure 15 reduced to just the top-voxel selection setting. I.e., in this comparison, all feature maps had euqual size, for all regions we selected the same amount of voxels (50). In particular LP-based measures are expected to be 'biased' to give higher scores for larger predictor (source) feature map sizes, as a special case of giving higher scores for higher-information-containing predictor feature maps. However, on the two fMRI datasets we tested, surprisingly, we observe no difference in performance of region matching in the two settings (compare fig. 15. Possibly, two factors offset each other: heightened region-matching capability by dropping the confounding factor of feature map size, and lowered region-matching capability by overall less information in the predictor features (50 is smaller than the typical number of voxels or vertices in each region.) This would

REPLICATION OF RESULTS ON A DIFFERENT SUBSET

To test how much our aggregate accuracy results depend on the exact set of regions selected, we select manually a second checkerboard-patterned subset, now not optimizing for high Glasser-atlas based anatomical reliability, but again ensuring a sufficient number of voxels have an NCSNR of 0.2 or above. This subset and its anatomic reliability and confusion matrix are shown in fig. 38. Note: For NSD, we drop region PFcm.

REGION-CONFUSION MATRICES FOR BASE MEASURES

BOLD MOMENTS

STATISTICAL METHODS - ADDITIONAL DETAILS

CONFIDENCE INTERVALS

Cohen's d_z
Following (Cumming & Finch, 2001), section "CIs for Cohen's $[d_z]$", we compute analytical confidence intervals for the (unbiased) parameter that d_z estimates - μ_d/σ_d, the ratio of the true mean difference μ_d to the true standard deviation of differences σ_d. We do this by solving for the noncentrality parameters Δ_L so that observing a t-value of at least the observed value has a probability of 0.025 ($(1 - level)/2$, where $level$ is 0.95 for 95% confidence intervals) , and symmetrically for Δ_U, and dividing those lower and upper bounds by $\sqrt{N_{subj}}$ to obtain lower and upper bounds for d_z. These CIs are shown in fig. 46 along with those for region matching accuracy.

Region-matching accuracy
We use Wilson score confidence intervals for Bernoulli variables.

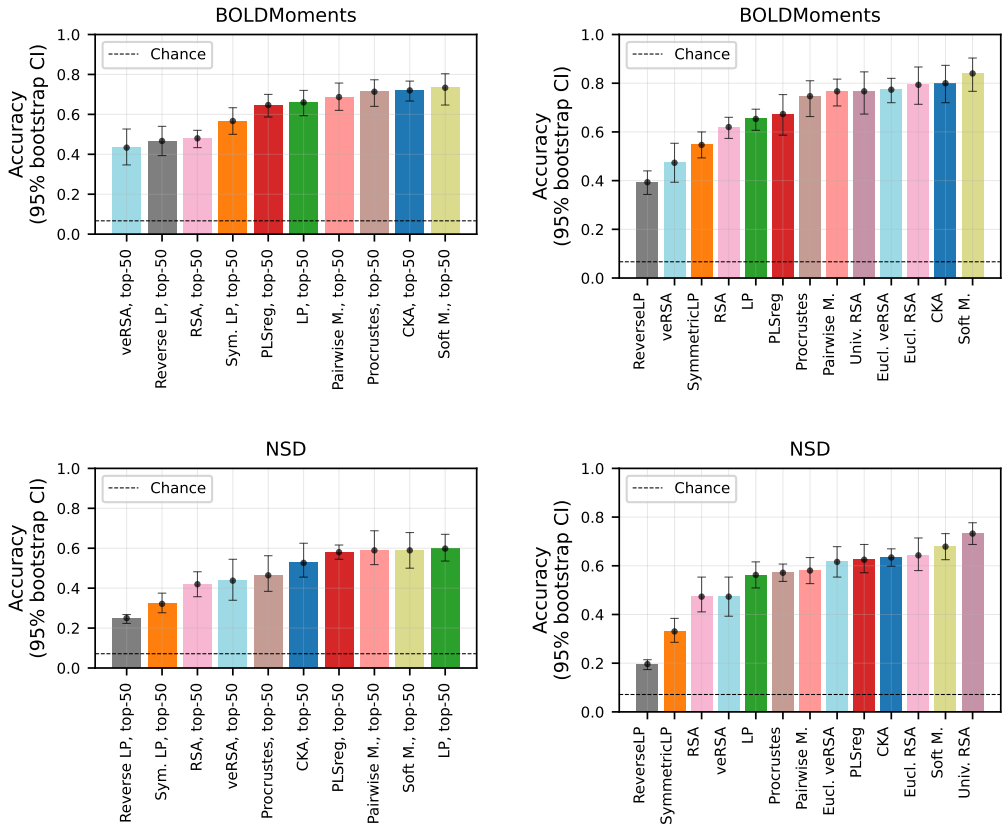


Figure 37: **Left:** Region-matching under equal feature map size: For each region, we select the top 50 voxels in terms of NCSNR, individually per subject, dropping subject-region pairs where no 50 voxels are above an NCSNR of 0.2. Note that top-50 voxel selection settings do not include results for Euclidean veRSA and Euclidean RSA. **Right:** For comparison, results under the default voxel selection procedure.

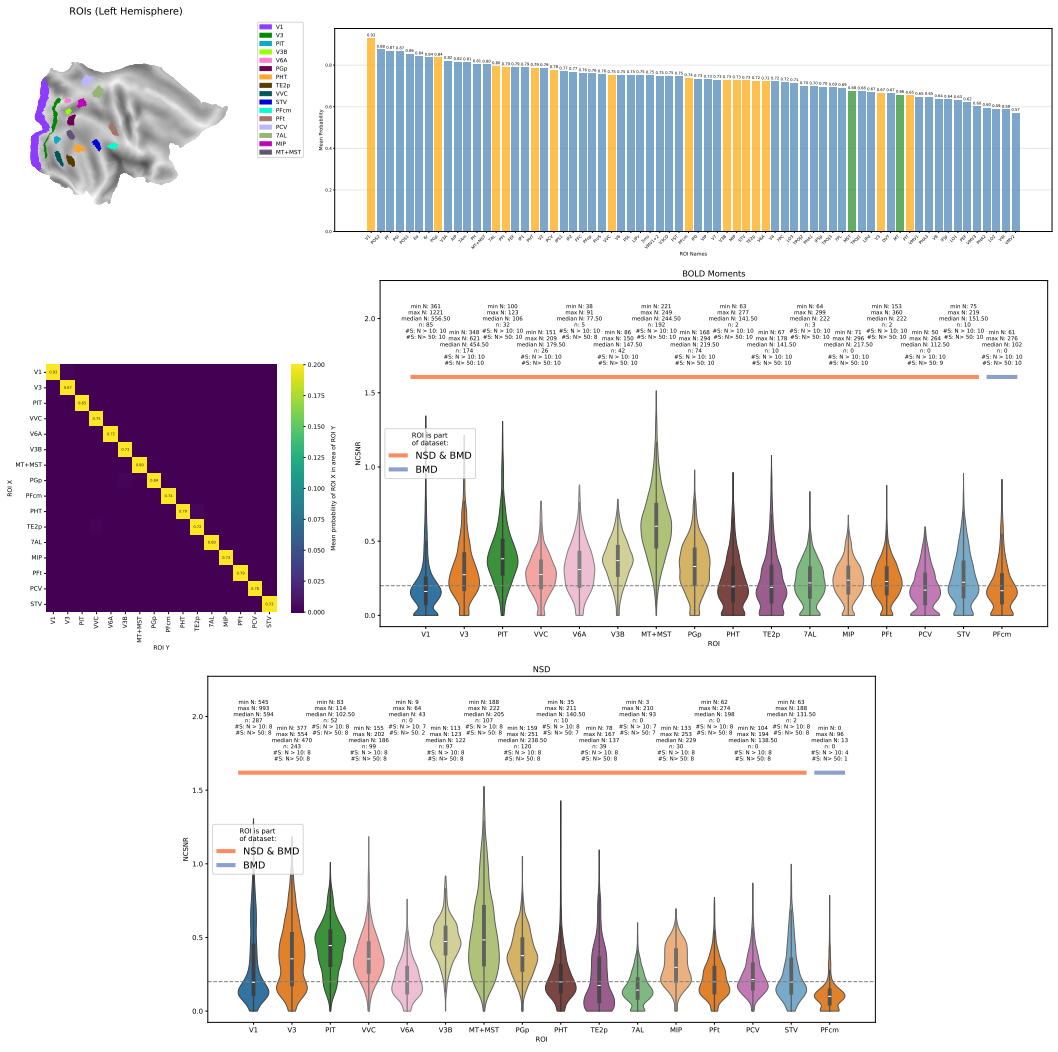


Figure 38: Distinct validation subset, again selected to exclude adjacent regions. **Top left**: The selected regions. **Top right**: Mean probability of the given region among the voxels within it, based on the Glasser-atlas validation set of subjects. **Center left**: Glasser-atlas probability map based overlap. **Center right:** NCSNR distribution for the regions of this set, for BMD. **Bottom:** Same for NSD. Note that here, not all regions satisfy that the minimum number of voxels above NCSNR of 0.2 is at least 10. Subject-region pairs not satisfying this condition are excluded from the analyses. In particular, while for all regions, at least seven subjects have more than 10 voxels above threshold, in V6A on NSD, only two subjects have at least 50 voxels surviving thresholding.

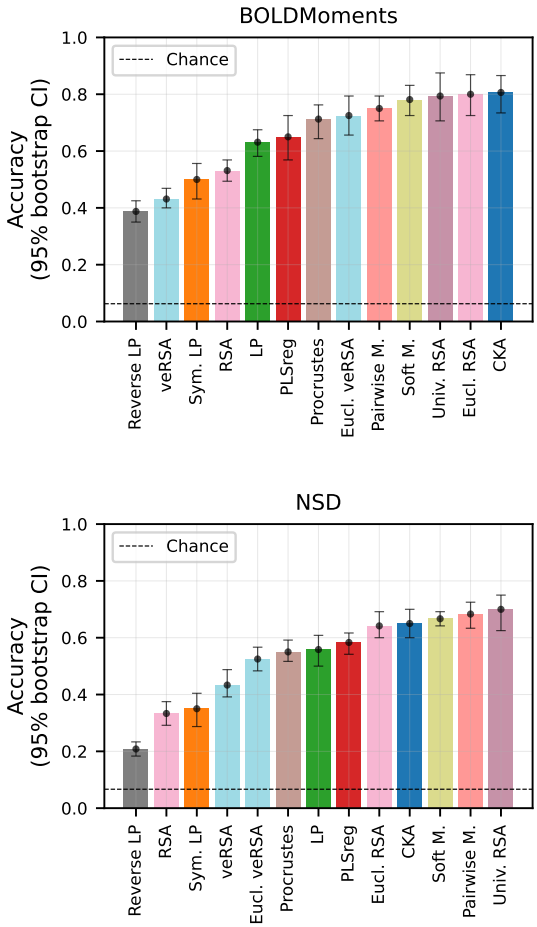
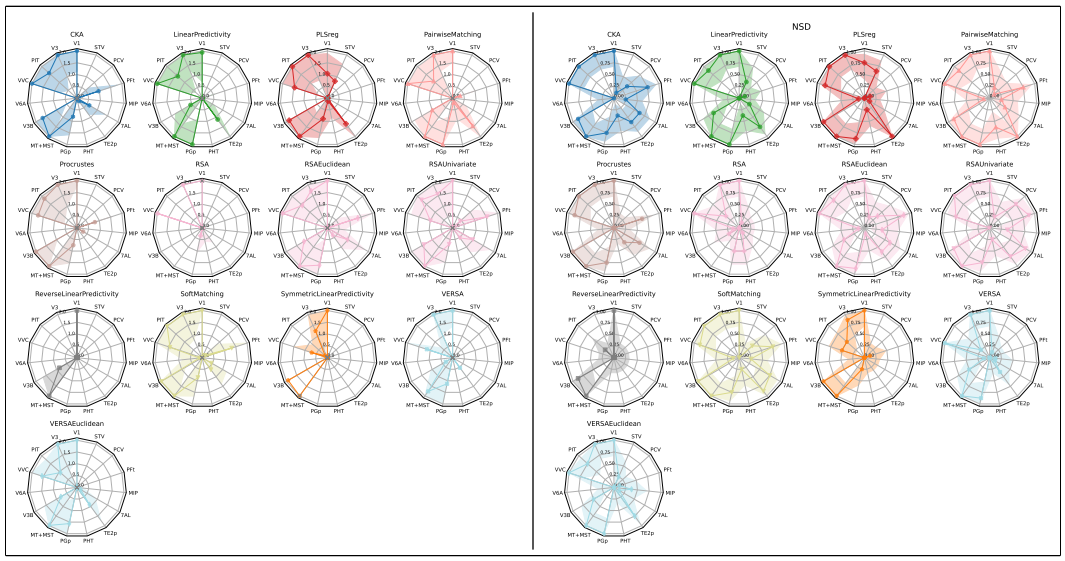


Figure 39: Region matching accuracy, validation set of regions (shown in fig. 38)

NSD

Cohen's d_z ROI-Matching Accuracy



BOLD Moments

Cohen's d_z ROI-Matching Accuracy

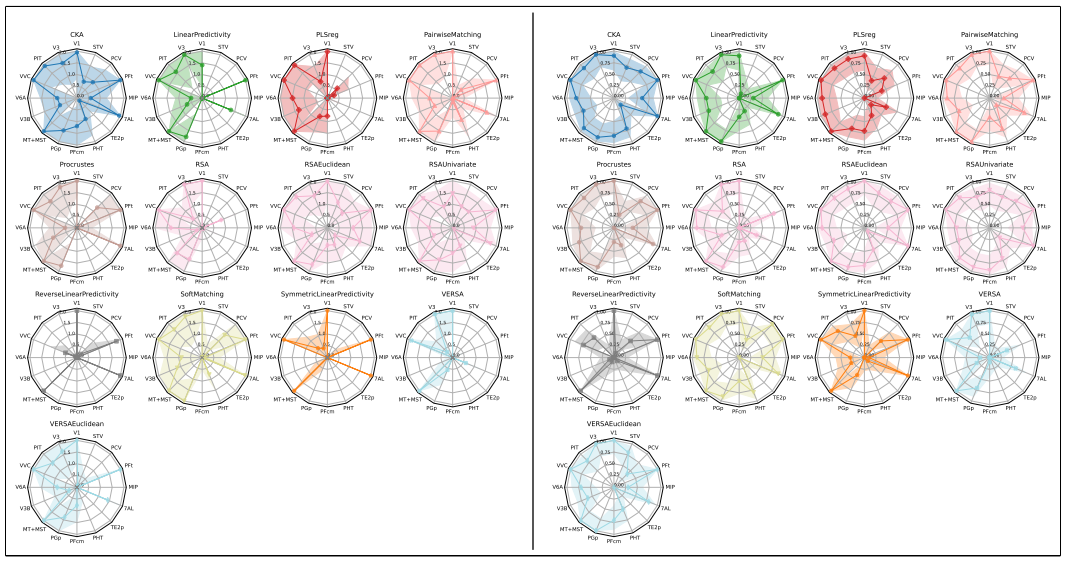


Figure 40: Measure d_z (left) and region-matching accuracy (right), with confidence intervals (CIs). d_z value range is clipped to $[0, 2]$. For our **validation set of regions** (shown in fig. 38).

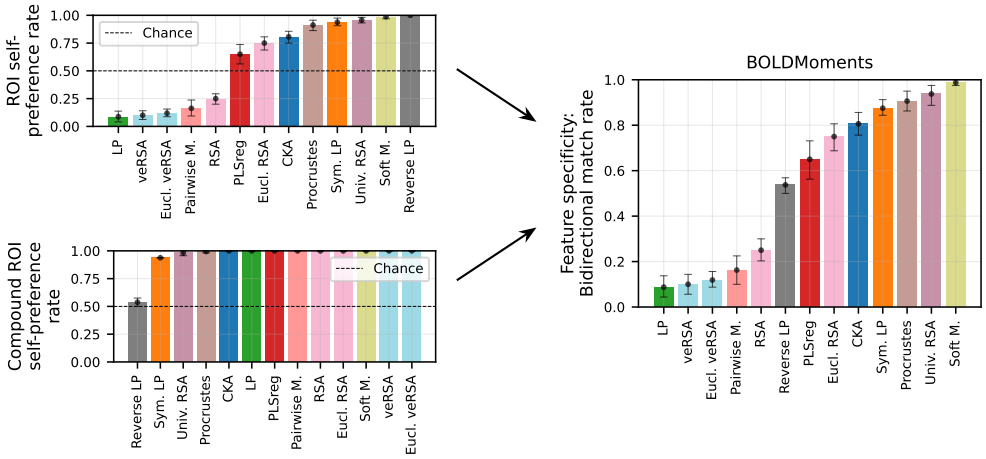


Figure 41: **Distinguishing regions from a compound region that contains them, for our validation set of regions (shown in fig. 38). BMD.** See also section 3.2.

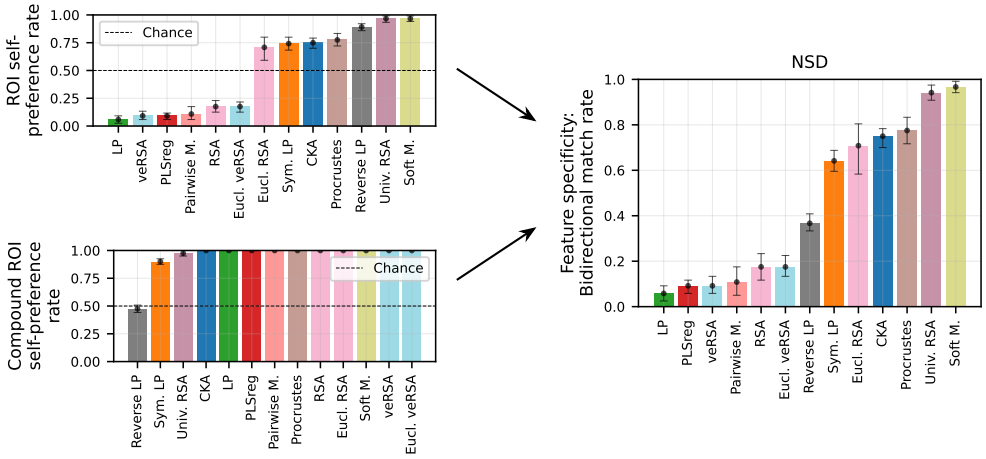


Figure 42: **Distinguishing regions from a compound region that contains them, for our validation set of regions (shown in fig. 38). NSD.** See also section 3.2.

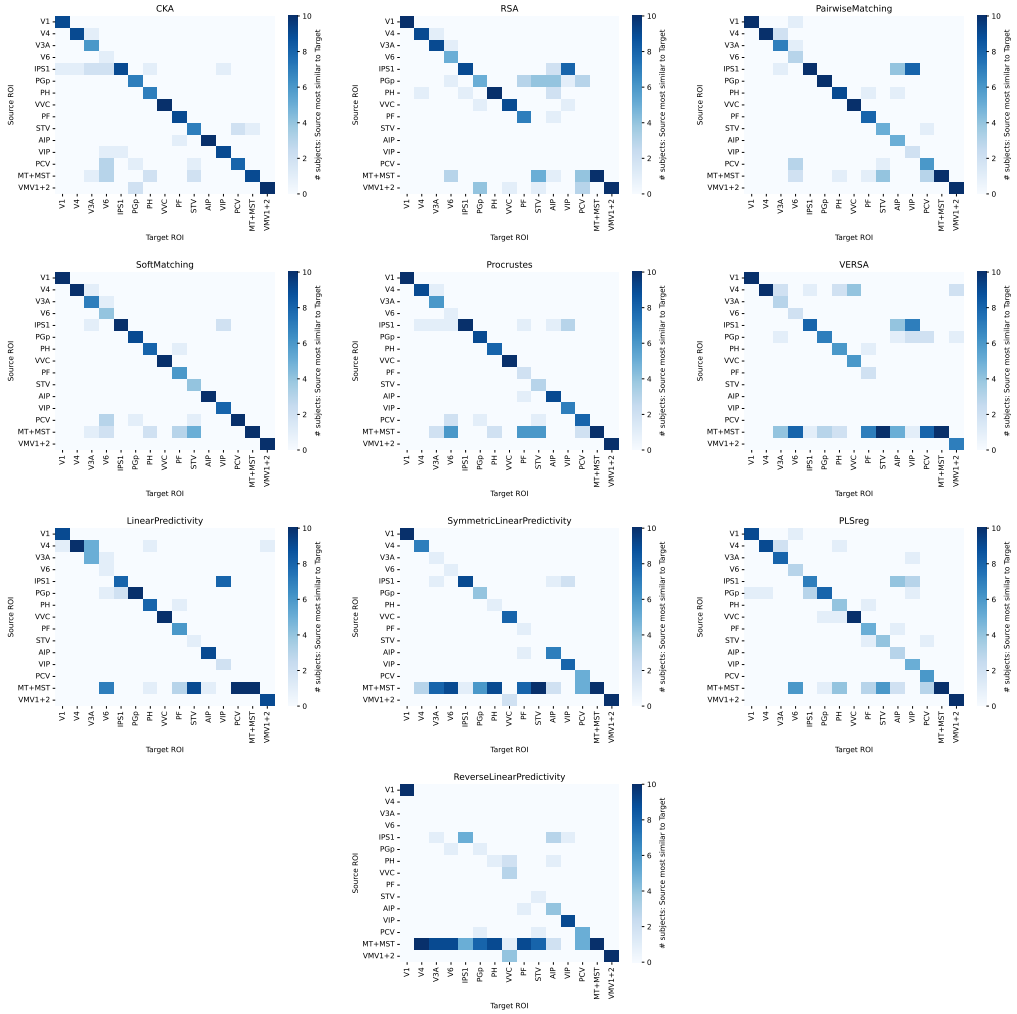


Figure 43: Confusion matrices on BMD for all measures. Scale goes from 0 (for no subject, those two regions were confused) to 10 (for all subjects, region i was the most similar region to region j). Asymmetry also for symmetric measures stems from the differences in source vs. target feature maps, where the former is a mean fMRI feature map across all except one subject, and the latter is the fMRI feature map for the remaining subject. Notwithstanding anatomical shifts, V2 is not confused with, nor a false match for, region VMV1+2. For SoftMatching and PairwiseMatching, it is a false positive for V6. It does present a considerable source of confusion for V1, most likely also because of functional proximity. Note that RSA and veRSA are Pearson-correlation based variants.

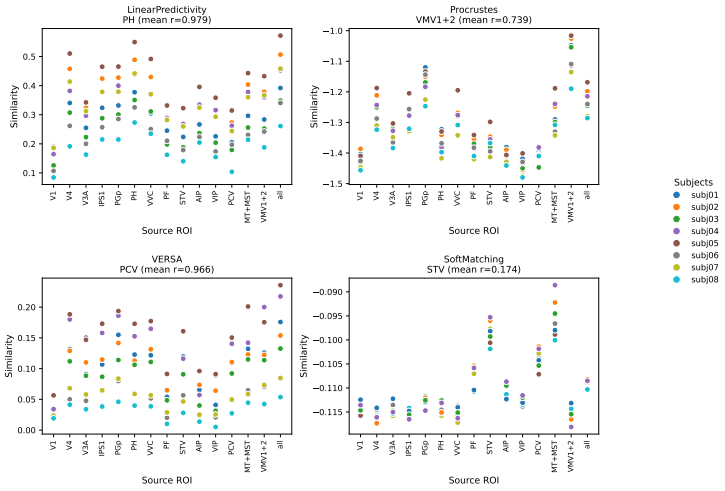


Figure 44: Similarity values colored by subject, for NSD. Right column: Two (similarity measure, target region) pairs sampled from the top-5% percentile of most strongly correlated similarity values across subjects (see text, K.3). Right column: two such pairs sampled at random from all subjects and regions we analyze.

Bias correction of Cohen's d and d_z

Both quantities are biased (Hedges, 1981), bias correction can be applied by multiplying with Hedge's correction factor ($\gamma = 1 - \frac{3}{(4\cdot(n-1)-1)}$ for d_z, which is approximately 0.9 for $n = 10$ subjects). This factor is constant across experiments on either of our datasets and therefore does not change metric order; omitting it leads to slightly inflated d_z values. For simplicity we report d_z values here, corresponding bias-corrected values would be scaled down by a factor of ca. 0.9

Why pairwise comparisons?

When there are considerable subject-specific effects in the data, this leads to additional variance that can overshadow true effects, if it is not accounted for. For example this can arise when a subject has particularly low noise and therefore its linear predictivity scores are higher than what would be expected based on other subjects. Figure 44 shows similarity values for a given target region and measure, colored by subject. The left column shows examples for strong subject-specific effects, i.e., 'base-levels' of similarity that differ by subject. There, differences between subjects often exceed differences between regions. Figure 45 quantifies this for both datasets, by showing the distribution of correlations between self-similarity (eg points for x-label PH in the scatter plot for region PH) and similarity between PH and another ROI. Correlations tend to be larger than 0.5 and clearly differ from the simulated null distributions shown in the bottom row.

Pairwise statistics deal with such effects by comparing similarities only within a pair. This effectively means that subject-specific additive offsets are ignored.

The silhouette score

Designed to measure cluster fit of a point i, the silhouette score (Rousseeuw, 1987) subtracts mean similarity to points in the next closest cluster (closest distractor ROI), $b(i)$, from the mean within-cluster (mean within-ROI) similarity $a(i)$ and normalizes by the larger one of the two:

$$s(i) = \frac{a(i) - b(i)}{max(b(i), a(i))}.$$

In our setting, a point could be a subject s and target region j: $i \overset{here}{=}$ (subject s, region j)). $b(s, j)$ would likely be the mean similarity between this subjects feature map for region j, and all

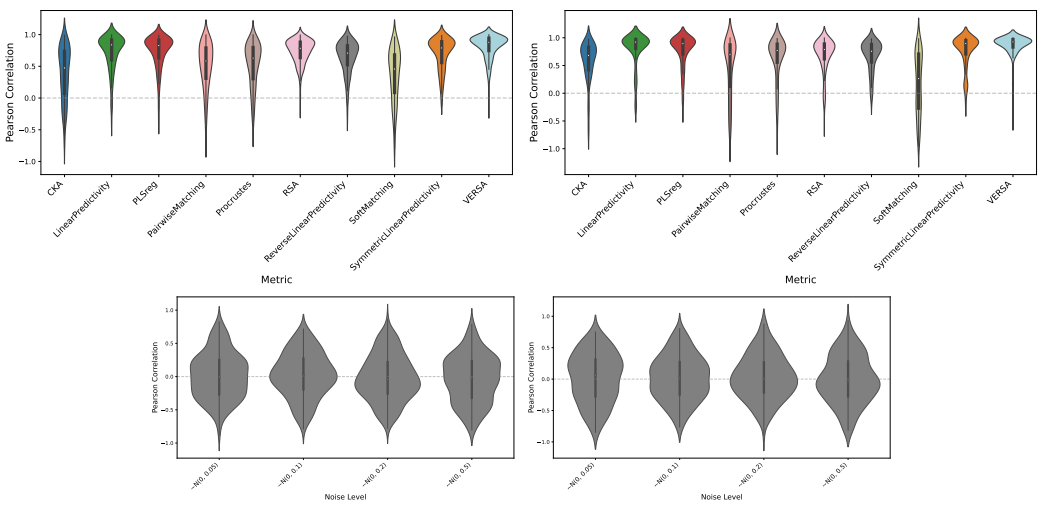


Figure 45: Empirical correlations ρ between vector pairs **??** and **??**, for all possible ROI pairs in our standard set of ROIs, for BMD (left) and NSD (right). Please see section K.3 for context. **Bottom:** The same statistic for simulated similarity values that are normally distributed (4 noise levels shown) and have trends across ROIs but not across subjects, for the same number of subjects and ROIs as in BMD (left) and NSD (right).

individual subjects feature maps, for the distractor region k for which this quantity is maximized. The original version of the silhouette score assumes symmetry of scores, but since it is defined for a given (s, j), point we can choose to always use this point as second argument (i.e., as target). $a(s, j)$ would then be the mean similarity of region j of subject s to region j in the other subjects.

The silhouette score assumes that similarity values follow a ratio scale (Rousseeuw, 1987) so that values are positive, and the zero point and ratios are meaningful. In our context, this means that ratios of scores must have the same meaning across measures - requiring all measures to, for example, report correlation-based R^2 values, if we consider R^2 values as lying on a ratio scale. See the difference between RSA and "squared RSA" in (Thobani et al., 2025). It has further disadvantages in our setting as being an unpaired method. Because of this and to allow any type of scale for similarity values, we believe that Cohen's d_z and accuracy allow easier comparison between measures and are therefore more suited for the type of analysis we propose.

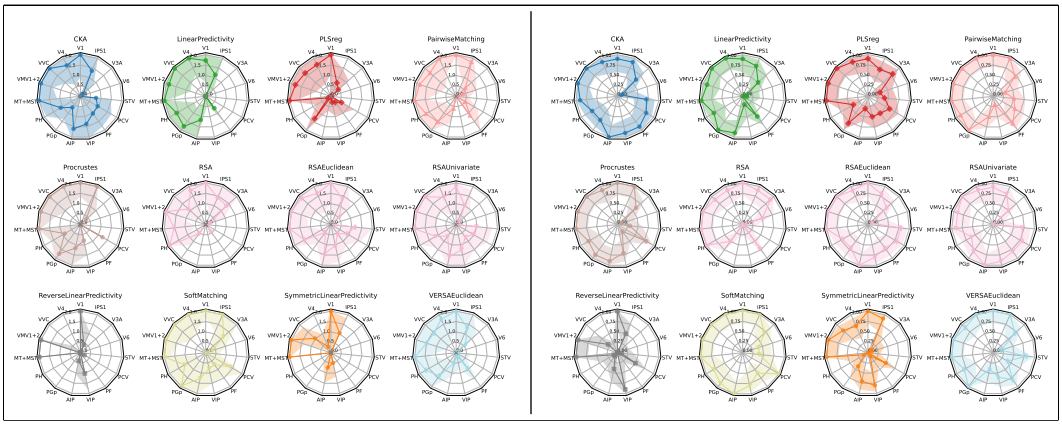
Comparing Cohen's d_z and region matching accuracy

We use two types of statistics to estimate how well a measure can distinguish a region from other regions: region-matching accuracy and Cohen's d_z. Do we have a similar issue here as with alignment measures, that we don't know which one to trust? Occasionally yes, they will disagree, largely they coincide: Figure 46 shows the two metrics side-by-side. Generally, curves looks similar for d_z and accuracy. If we reject the use of a measure based on low accuracy or d_z, it depends on the thresholds we set whether the two approaches agree. As shown in figure 47, for each accuracy threshold there is a dz threshold that leads to acceptance of the same measures for the same ROIs in around 95% of cases (8 of 150 pairs, BMD) or 97% (4 of 140, NSD). As a comparison between the two quantities, we can test, for example, how frequently (for how many region-measure pairs) the lower end of a confidence interval is at or below zero. On BMD, is the case in 14 out of 150 pairs for accuracy (9%) and in 41 out of 123 pairs (33%; this excludes CIs where computation failed) for d_z. For NSD, for accuracy this is the case in 37/140 pairs (26%) and for d_z, in 32/91 pairs (35%).

Leave-one-out mean leads to higher similarity scores compared to pairwise similarities

BOLD Moments

Cohen's d_z ROI-Matching Accuracy



NSD

Cohen's d_z ROI-Matching Accuracy

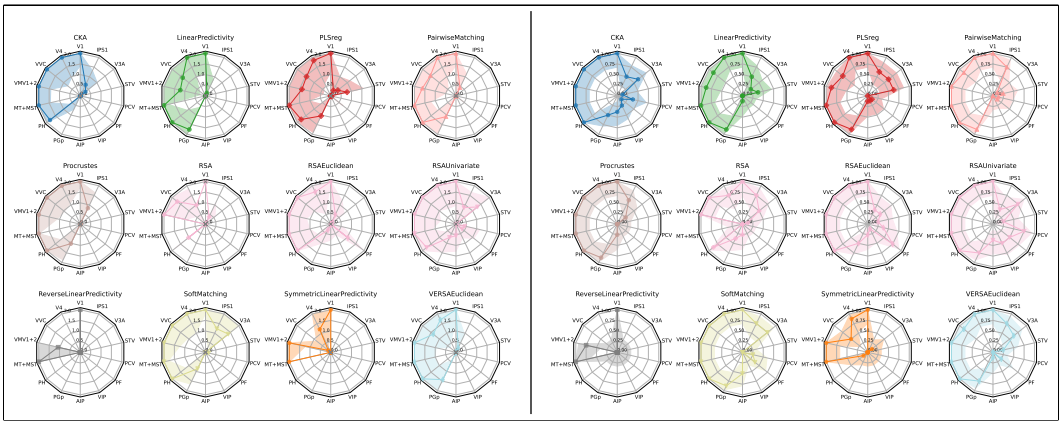


Figure 46: Measure d_z (left) and region-matching accuracy (right), with confidence intervals (CIs). d_z value range is clipped to $[0, 2]$ so that a marker at value 2 indicates a d_z value of 2 *or above*. For d_z, CI computation can fail. A missing CI is indicated by a small gray cross on top of a marker. (E.g.: d_z on NSD in V1 for RSA.) Note that CIs can also appear missing for d_z if the CI completely falls out of the shown range of $[0, 2]$. CIs are on the level of 95%. CI's are Wilson score intervals for accuracy values (right column), and CIs for d_z are computed as described in appendix section K.1.

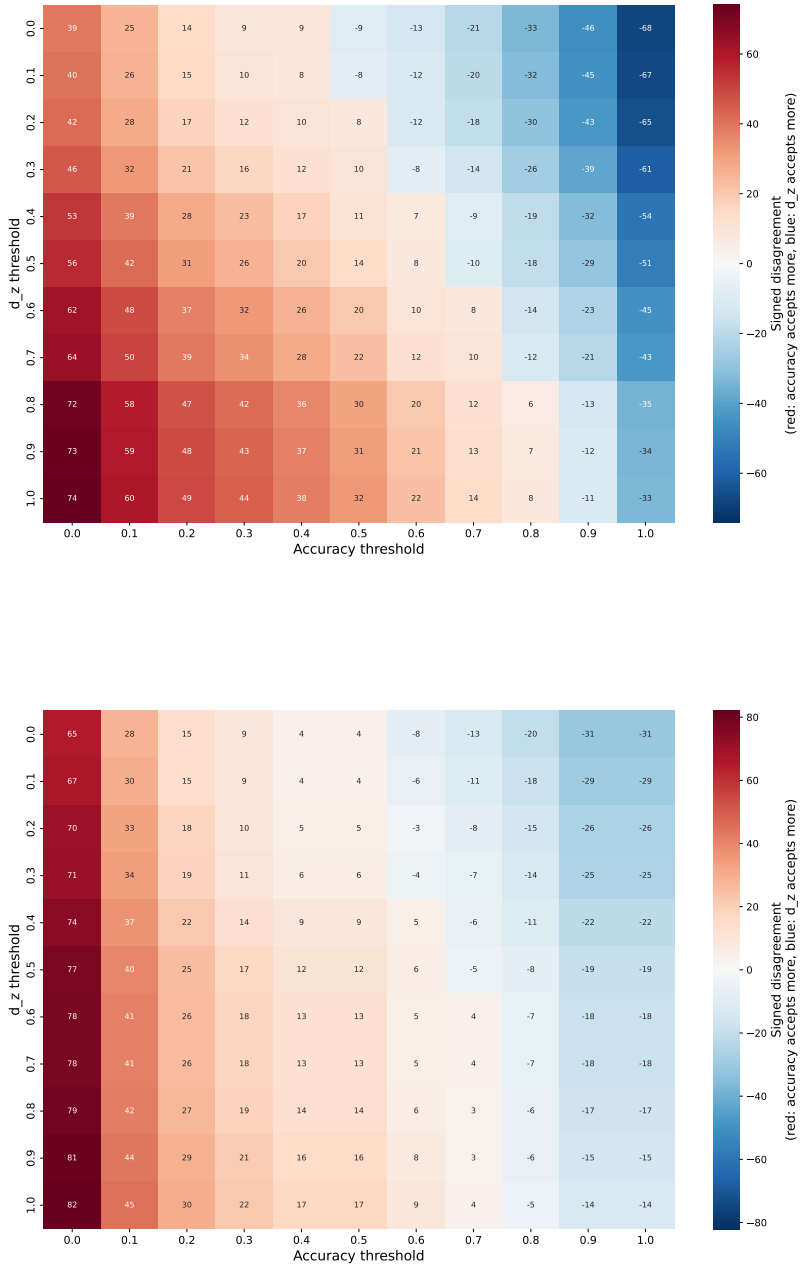


Figure 47: Absolute disagreement when using either a threshold on region-matching accuracy (x-axis) or Cohen's d_z (y-axis) as criterion to accept use of a measure for a given ROI. Shown are the number of pairs of measure and target ROI, out of 150 total pairs (140 for NSD), where the two thresholding methods disagreed whether the measure fails the sanity check or not. **Top:** BMD. **Bottom:** NSD. Rois are the 15 and 14 ROIs for BMD and NSD, respectively, measures include all standard versions of the measures listed in section C *except* that this does not yet include Euclidean versions of RSA and veRSA, but Pearson-RSA and Pearson-veRSA.

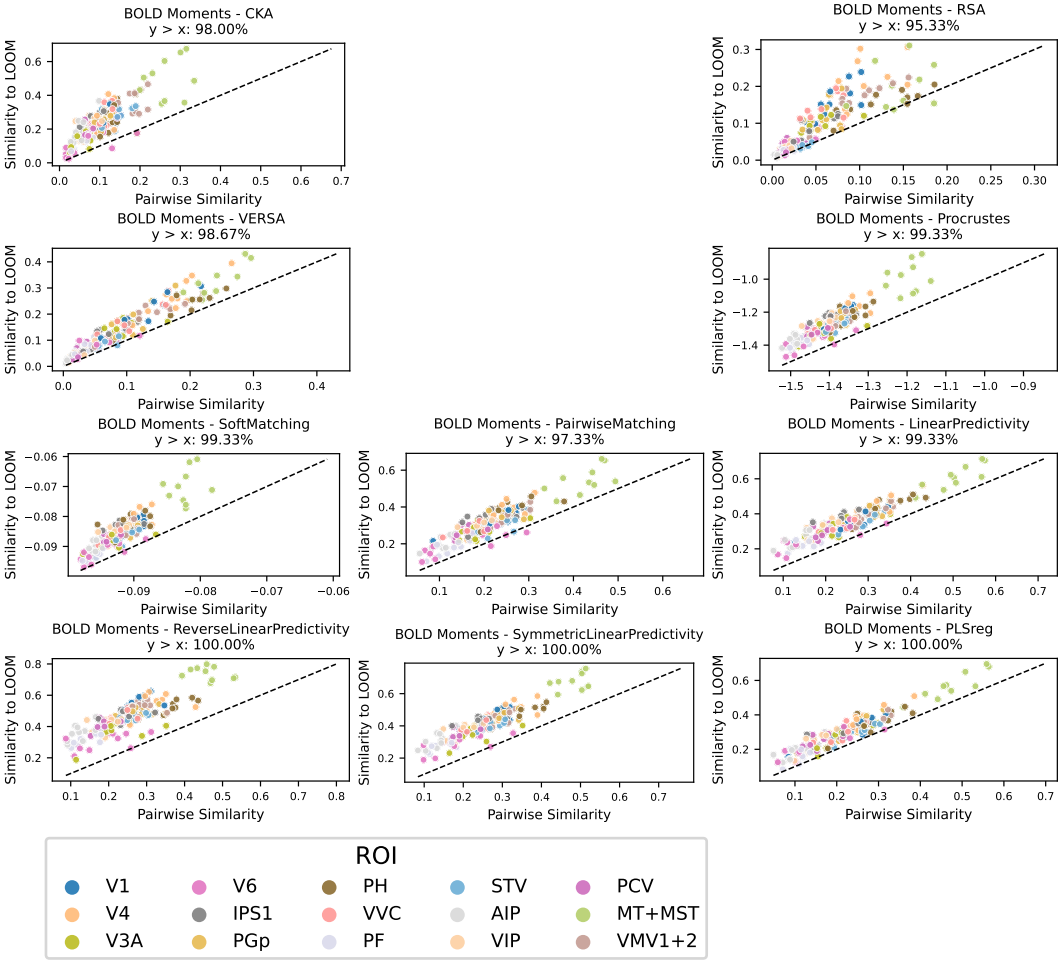


Figure 48: **What is a good ground truth model? Dataset: BMD.** Similarities for 'correct matches', computed against leave-one-subject-out mean (*LOOM*) fMRI features as 'model', vs. individual subjects' fMRI features as 'model'. Each point represents the comparison between one ROI of the 'model', to the same ROI of one target subject. Its y axis location is the similarity from using leave-one-subject-out mean fMRI features as 'model', its x-axis location represents the similarity from using another randomly assigned subjects' feature map as 'model'.
Almost always, *LOOM* features of the correct ROI score higher similarities than single-subject features of the same ROI.

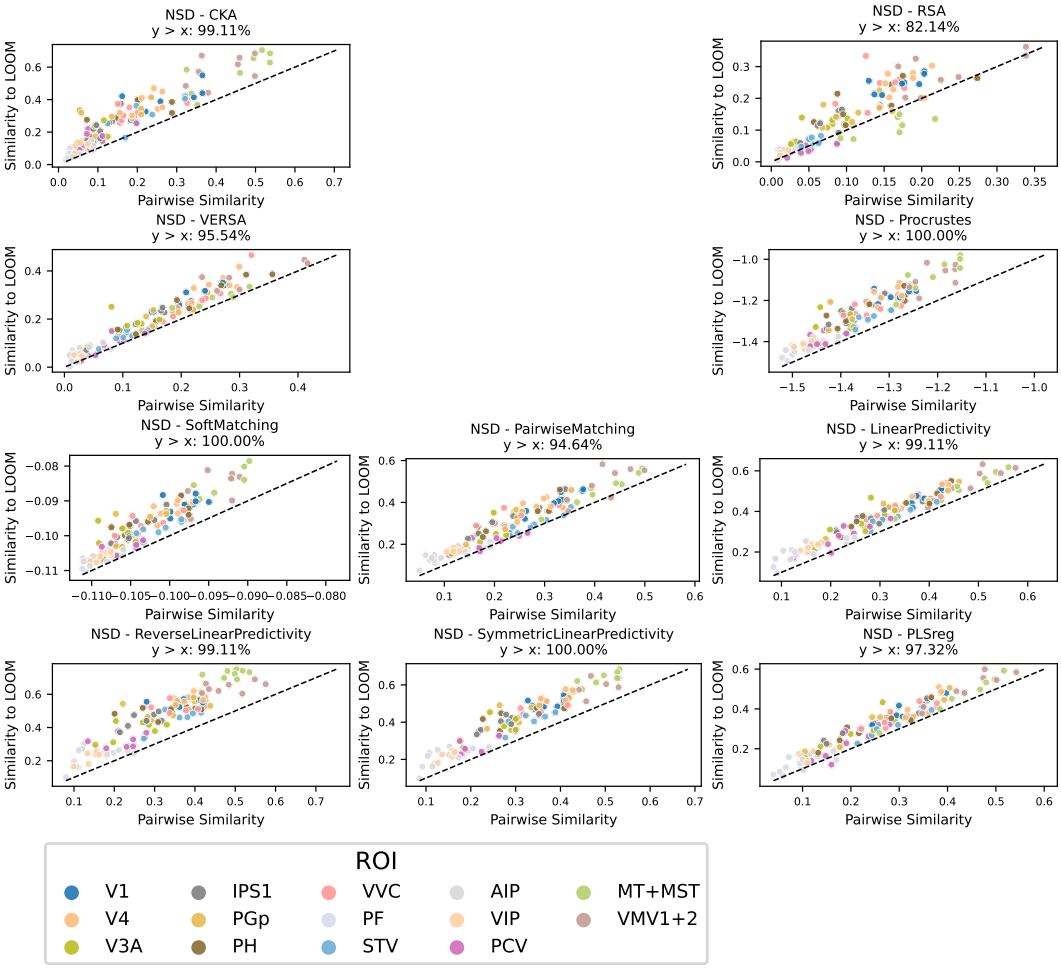


Figure 49: **What is a good ground truth model? Dataset: NSD.** Same as figure 48, for NSD.