Computational Brain & Behavior
https://doi.org/10.1007/s42113-025-00257-5

RESEARCH

How Robust Are fMRI- and EEG-Based Representational Similarity Analysis?

Satwick Sen Sarma[1] · **Gouravmoy Boruah**[1] · **Nisheeth Srivastava**[2]

Accepted: 25 July 2025

Abstract
In EEG and fMRI analysis, researchers choose from a combinatorially large set of theoretically indistinguishable options while building a data processing pipeline based on individual beliefs and other factors. However, not all pipelines are reliable, although the same might not be evident while performing the analysis. Thus, re-analyzing the data through various alternate pipeline configurations presents an opportunity to determine the reliability of the reported results. Results that are replicated across a larger number of pipeline specifications may be considered reliable. In this paper, we adapt a technique recently developed in the psychology literature, specification curve analysis (SCA), to quantitatively assess the robustness of noninvasive neuromodeling results drawn from fMRI and EEG data. Our empirical results, based on a reanalysis of the THINGS dataset, show that the conclusions drawn from EEG-based RSA are fairly robust to alternative specifications, but not fMRI-based RSA. We present a decision tree-based approach to identifying the most robust specification given a set of alternative specifications and dataset for any analysis. However, even the most robust set of specifications for fMRI-based analysis still yield fairly epistemically fragile conclusions. We conclude, based on these results, that SCA could and should be applied without loss of generality to nearly all event-related fMRI data analysis protocols, with suitable modifications to the set of alternative specifications we used in our work.

Keywords Specification curve analysis · fMRI analysis · Representational similarity analysis · Deep neural network

Introduction

In computational neuromodeling, pipeline configurations constitute an extremely selective set of analytical choices and, as a result, rarely map bijectively to underlying assumptions. For the same neural computational model assumptions, there exist multiple combinations of valid parameter choices for the encoding and decoding of the model estimation pipelines. The chances of introducing confounding results in certain combinations threaten the reliability of the model estimates and the downstream analysis (Kelly Jr & Hoptman, 2022). Evaluating the consistency of estimates across specifications is necessary for inferences resulting from processing pipelines with high degrees of freedom.

For example, multivariate encoding and decoding analyses in fMRI and EEG, respectively, work under a similar hypothesis of linearized feature space to transform feature representations from the feature space to the activity space, which forms the basis for the development of linear encoding and decoding models (Naselaris et al., 2011). Usually, multiple combinations of specifications are supported by the same assumptions as true computational models. Likewise, pipeline specifications span from this singular latent assumption to multidimensional parameter space for both modalities, and more so for fMRI pipelines. For example, to estimate feature vectors based on linear encoding models (Naselaris et al., 2011) from BOLD fMRI, the estimation of the best-fitting canonical hemodynamic response function can be performed

Satwick Sen Sarma and Gouravmoy Boruah contributed equally to this work.

✉ Nisheeth Srivastava
nisheeths@gmail.com

Satwick Sen Sarma
satwick22@iitk.ac.in

Gouravmoy Boruah
gouravmoy22@iitk.ac.in

[1] Department of Cognitive Science, IIT Kanpur, Kalyanpur, Kanpur, Uttar Pradesh 208016, India

[2] Department of Computer Science and Engineering, IIT Kanpur, Kalyanpur, Kanpur, Uttar Pradesh 208016, India

Published online: 14 August 2025

in a multitude of ways (Pedregosa et al., 2015). Moreover, estimation strategies for the same, such as voxel-wise best-fitting HRF, often introduce confounds due to idiosyncratic noise embedded in the signal in individual voxels.

In contrast, the alternative choice of estimating HRFs for regions of interest (ROI) results in more unbiased estimates. Generally, such analysis forks are present in processing pipelines in neuromodeling analyses at each parameter choice. For a limited sample size in modalities like fMRI (Turner et al., 2018), these potential biases amplify the concerns of Type-1 and Type-II errors. Inferences based on a specific pipeline configuration might not generalize across populations or, worse, across alternate pipeline configurations.

In contrast, EEG offers a unique set of challenges and opportunities when it comes to decoding neural activity. Unlike fMRI, it offers high temporal resolution and direct measurement of population-level neural activity (Cohen, 2017), aiding in a detailed examination of the temporal aspects of representational dynamics. Its constraints, such as spatial resolution and signal-to-noise ratio, are offset by its temporal precision, crucial for effective decoding analysis. However, EEG analysis is complicated by factors such as its inherent non-linearity and non-stationary properties (Gramfort et al., 2013; Cole & Voytek, 2019), as well as interindividual variability (Clerc et al., 2016), all of which can significantly affect the analysis and decoding performance.

In a similar vein, decoding models are often limited to classifying between only a few data conditions, using training sequences derived from the same conditions (Haynes & Rees, 2006). The limited scope of these models thus affects their ability to generalize, as they often struggle to adapt to a wide variety of brain states, leading to a propensity for classifier overfitting. Multivariate EEG decoding analysis often uses classifiers that allow them to detect subtle patterns missed in average signal comparisons, typically used in univariate methods (Grootswagers et al., 2017); however, classifiers are sensitive and require caution when interpreting decoding results (Pereira et al., 2009); checks must be in place to ensure that classifiers learn categorical distinctions that help them generalize, rather than learning specific exemplar features (Carlson et al., 2013). However, the decoding pipelines used in most EEG studies are rarely tested for robustness, similar to the practices of fMRI studies, and therefore should also be subjected to testing for alternative specifications.

These concerns about robustness converge, by design, on representational similarity analysis (RSA), a recent adaptation of classical MVPA, which facilitates the comparison of representational dynamics between modalities, primarily between brain activity, behavioral response, and computational models (Kriegeskorte et al., 2008). In the last decade, there has been an efflorescence of studies using RSA to test strong and, at times, ambitious hypotheses. Representations under RSA are built at a higher level of description (Shea, 2018) with strong assumptions, viz, that brain or behavioral activity seen in an experimental condition can be directly treated as a stimulus-condition representation. Comparison of representational geometry is reduced to a correlation of response vector distances across the stimulus conditions in the respective modalities, encapsulated as entries in representational dissimilarity matrices (RDMs).

However, the data entering RSA are fraught with the same perils as any other neuromodeling pipeline, offering the researcher a plethora of alternatives for various analytical choices—from HRF estimation, choice of brain voxels for regions of interest, and a few added dimension for the RSA specific computations such as distance measure to comparison metrics. With the massive number of computations necessary for computing an RDM, the potential noise sources in estimations are a pressing concern first for the RDMs themselves and secondly for any downstream comparisons using them. Similar concerns have previously been raised (Ritchie et al., 2021), with suggestions for repeating the analyses using alternate pipeline configurations to evaluate the reliability of the estimates resulting from an RSA pipeline, specifically for a modality such as fMRI with a low signal-to-noise ratio (SNR).

In this paper, we adapt a technique recently developed in the psychology literature, specification curve analysis (SCA) (Simonsohn et al., 2020), to quantitatively assess the robustness of non-invasive neuromodeling results drawn from fMRI and EEG data. SCA is built upon a simple principle—researchers choose from a combinatorially large set of theoretically indistinguishable options while constructing a data processing pipeline based on individual beliefs and other factors. However, not all pipelines are reliable, although the same might not be evident while performing the analysis. Thus, re-analyzing the data through various alternate pipeline configurations presents an opportunity to determine the reliability of the reported results. Results that are replicated across a larger number of pipeline specifications may be considered reliable. SCA allows one to redo the analysis in these specifications, the original pipeline configuration being one of them, and report whether the result continues to be statistically significant in the most reasonable specifications of the analysis chain or whether a very specific set of choices leads to a statistically significant result, while most others do not (Simonsohn et al., 2020).

We additionally estimated the robustness of model-based hypothesis testing based on fMRI data using a similar SCA approach and present an algorithm for finding the most robust specification for an analysis pipeline, given a set of alternative specifications and a dataset.

Methods

Dataset

To estimate the reliability of the RSA based on task-based fMRI and EEG signals, we chose datasets that included multiple sessions for the same visual stimulus categories. First, for fMRI signals with low signal-to-noise ratio (SNR), the multi-session response for stimulus categories is essential for inference sensitive to the response characteristics of the categories, with the high degree of noise in event-related and sparse block-based designs. Secondly, training classifiers to perform decoding analysis in EEG requires a dataset with multiple observations for individual object categories. In addition to the dense individual sampling (Turner et al., 2018), evaluating the reliability using a dataset encompassing a diverse set of stimulus categories also ensures the generalizability of the inferences from SCA to broader domains of studies. To this end, our initial analyses for fMRI and EEG used THINGS (Hebart et al., 2023; Grootswagers et al., 2022) datasets of the respective modalities curated specifically for representational similarity analysis with the design structure desired for our analysis. For the subsequent fMRI analysis to assess the reliability of model hypotheses using RSA, we performed that analysis using a dataset with fewer categories but denser sampling per individual, i.e., more sessions for each image category. In addition, we chose a dataset that has previously been tested for similar model hypotheses. For this purpose, we used this dataset (Horikawa & Kamitani, 2017) that has been previously tested (Choksi et al., 2022) for model hypotheses.

fMRI Data

For our primary assessment of the reliability of RSA, we used this dataset (Hebart et al., 2023). THINGS fMRI dataset consisted of data from three subjects. For each subject, 12 sessions of the BOLD response of a diverse set of 720 representative object categories with a different exemplar from each category were recorded in the sessions. In addition to the category responses, the dataset included 100 test images and unique synthetically generated catch images in all sessions. For our analysis, we only used the functional data recorded in response to the category exemplars of 720 diverse object concepts. Functional magnetic resonance data were collected using three Tesla Siemens Magnetom Prisma scanners with a 32-channel head coil. The fMRI task, functional brain MRI data was recorded with isotropic resolution of 2 mm (60 axial slices, 2 mm slice thickness, no slice gap, matrix size 96×96, FOV = 192×192 mm, TR = 1.5 s, TE = 33 ms, flip angle = 75°, echo spacing 0.55 ms, bandwidth 2,264 Hz / pixel, multiband slice acceleration factor 3, phase encoding posterior to anterior) (for further methodological details, see Hebart et al. (2023)). The stimulus was presented in an event-related design for 500 ms, followed by a fixation of 4 s. In addition to the functional data, six categories of selective functional runs for faces, body parts, scenes, words, and objects were recorded for each subject. We used category localization runs to estimate region masks. For our analysis, we focussed on the extrastriate body area (EBA), the parahippocampal place area (PPA), the fusiform face area (FFA), and the lateral occipital complex (LOC). For each region, we estimate the mask using the categories-selective runs and the region atlas (Julian et al., 2012). Since the RDM estimates for individuals are independent of other subjects, for computational tractability, we performed the analysis on a randomly selected subject from the pool of three subjects.

To test the robustness of the model hypotheses in response to analytical flexibility, we used (Horikawa & Kamitani, 2017) composed of recordings in response to a subset of ImageNet image categories. The dataset included a BOLD response of 1200 training images for five subjects. In addition, functional data were recorded in response to 50 test images in 35 sessions for each individual. Given that the objective of our analysis was to estimate the interaction between the analysis decision and the significance of the hypothesis, the average response to the repeated presentation of 50 test images provided the optimal basis for the creation of RDM to decrease the effect of noise at an intra-subject level. Functional fMRI data were collected using a 3.0-Tesla Siemens MAGNETOM Trio A Tim scanner located in the ATR Brain Activity Imaging Center with no cut gap, matrix size 64x64, slice thickness 3 mm, FOV = 192x192mm, TR = 3ms, TE = 30ms, flip angle = 80° (for further methodological details, see Horikawa and Kamitani, 2017).

EEG Data

For EEG, we used the THINGS-EEG dataset (Grootswagers et al., 2022) for our RSA reliability assessment. The data used were from five subjects randomly pooled from a sample size of 50 subjects. The stimuli shown during the experiment were from the THINGS database (Hebart et al., 2019). The analysis used data from the 200 validation images, which were shown in random order using a rapid serial visual presentation (RSVP) paradigm, repeated in 12 sequences. Continuous data were recorded using a 64-electrode BrainVision ActiChamp system, arranged according to the international 10-10 system standard for electrode placement. The digitization of the signal occurred at a resolution of 0.0488281 microvolts, with a sampling rate of 1000 Hz, and the electrodes were referenced online to Cz. Each image was presented for 50 ms, followed by a blank screen lasting another 50 ms. The same 200 images were consistently shown to all subjects, as the uniformity of the stimuli

between participants is essential for accurately assessing the reliability of the underlying representational dynamics.

Preprocessing

fMRI

The functional magnetic resonance imaging data was preprocessed by performing slice timing correction, rigid head motion correction, field map-based susceptibility distortion correction, alignment of the functional space with the individual subject's T1-weighted anatomical template (co-registration was implemented with nine degrees of freedom), segmentation of brain tissue, and reconstruction of the surface of pial and white matter. All preprocessing steps were implemented using fMRIprep. Following the methodology for surface reconstruction in Hebart et al. (2023), surface reconstruction was performed using Freesurfer recon-all to use all available T1-weighted and T2-weighted anatomical images for each subject and subsequently passing the output to fMRIprep downstream preprocessing steps. However, for Horikawa and Kamitani (2017), due to anatomical scans containing only one instance of T1-weighted and T2-weighted images for each subject, the default module of Freesurfer's recon-all available in the fMRIprep workflow was utilized for surface reconstruction.

EEG

The EEG data was preprocessed offline using Matlab (R2020b) (Inc., 2020) and the EEGlab (v14.0.0b) toolbox (Delorme & Makeig, 2004). The continuous data was filtered using a Hamming-windowed FIR filter, using a 0.1 Hz high-pass filter to remove low-frequency noise and a 100 Hz low-pass filter to limit high-frequency content. The data from the electrodes was then re-referenced to the average reference and then downsampled to a sampling rate of 250 Hz. The continuous EEG data were then epoched into trials ranging from 100 ms before to 1000 ms after stimulus onset, with each epoch containing data from 275 time points.

fMRI Single-Trial Response Estimates

Data Denoising

Estimates of single trial response were estimated in a stepwise manner. First, GLM was fitted to eliminate the noise components from the response patterns. For the THINGS dataset (Hebart et al., 2023), the noise component was estimated by replicating the methodology in Hebart et al. (2023). ICA was implemented using ICA-MELODIC after smoothing and high-pass filtering of the preprocessed signals, and noise components were detected using preidentified threshold configurations by human raters, as mentioned. For this dataset (Horikawa & Kamitani, 2017), the nuisance regressors generated by fMRIprep were used for noise normalization. Nuisance regressors for six basic motion parameters (three translational and three rotational), frame-wise displacement FD, and the first ten anatomical component corrections aCompCor based on the highest eigenvalues were used to denoising each functional run. In addition to these noise regressors, noise regressors are also constituted of polynomial drift regressors up to degree 4 for denoising both datasets. Regressing the noise components and the estimation of the response pattern was performed in a stepwise manner to attribute variance to the noise components independently of the stimulus responses and the higher noise ceiling estimates, even in the presence of collinearity between the components and the brain response (Kay et al., 2013).

General Linear Modeling

BOLD response amplitude was estimated by implementing the GLMs in a similar manner to the GLMsingle (Prince et al., 2022) but tuned for stepwise denoising and the estimation of the response amplitude repeated across sessions for Hebart et al. (2023) and runs for the datasets (Horikawa & Kamitani, 2017). For single-trial response estimates, the hemodynamic response function (HRF) was modeled using a library available from 20 available HRFs (Allen et al., 2022; Prince et al., 2022) to account for the variability of response patterns for individual voxels. The HRF that performed best for each voxel was estimated by devising multiple iterations of design matrices about each of the HRFs and identifying the HRF with the highest mean $R2$ in the stimulus presentations. Ridge regression models were used to estimate the amplitudes of the BOLD response. Hyperparameter tuning was performed to find the optimal value of the regularization parameter for each voxel using a comprehensive set of parameter values encompassing 0.1 to 0.9 in intervals of 0.1 and from 0 to 0.1 in step sizes of 0.01. The hyperparameter tuning revealed that 0.1 is the most optimal with low variability across voxels, and the regularization parameter was subsequently set to 0.1 for subsequent model fits. HRF best-fit estimation and hyperparameter tuning for regularization parameter were performed using leave-one-out cross-validation, respectively, for the 12 sessions of THINGS dataset using the 100 repeated images and 35 repetitions of 50 images in the dataset (Horikawa & Kamitani, 2017). However, for the latter, we performed a voxel-specific hemodynamic response function estimation in contrast to region-specific HRFs identified for the THINGS dataset. Although voxel-wise HRF identification is prone to more biased estimates (Badillo et al., 2013), due to the unavailability of individual region masks, we performed voxel-wise HRF identification. Finally, after the response estimation, to debias the beta coefficients from ridge regres-

sion, we estimated the final response amplitudes by linearly rescaling the regularized coefficients against the unregularized coefficients using a subsequent regression model fit.

For this dataset (Horikawa & Kamitani, 2017), the beta coefficients estimated in the individual T1w space were transformed into the MNI305 space using Freesurfer's mri-vol2vol (Fischl, 2012) for inferences between subjects.

Region Mask Estimation

Region selective masks were estimated using the six category localizer runs based on region-specific T contrasts for the selective response to object categories—body parts >objects (for EBA), faces >objects (for FFA), scenes >objects (for PPA), and objects >scrambled (for LOC). For estimating the statistical maps, the functional data were spatially smoothed (FWHM = 5 mm) and subsequently entered as regressors for each category, i.e., body parts, faces, objects, scenes, words, and scrambled objects. The resulting statistical parametric maps aggregated across functional runs with a fixed effects model (Woolrich et al., 2004) with corrections for multiple comparisons (cluster *p*- threshold=0.0001, extent-threshold=3.7). Finally, ROI masks were intersected with an existing group segmentation of category-selective masks (Julian et al., 2012) to generate region-specific masks for the fusiform face area (FFA), the occipital face area (OFA), the extrastriate body area (EBA), the parahippocampal place area (PPA), and the lateral occipital cortex (LOC).

Since the dataset (Horikawa & Kamitani, 2017) did not contain any category-specific runs, region-specific masks were estimated using the Desikian-Killiany atlas in the MNI305 template. To perform the analysis on this dataset, we identified the top 30 voxels for each region based on the magnitude of the response using the same methodology as Choksi et al. (2022) and included the same as one of the analysis forks in the specifications tree.

EEG Multivariate Pattern Analysis

For the THINGS-EEG dataset (Grootswagers et al., 2022), the EEG responses evoked for 200 images for five subjects were used for our multivariate pattern analysis (Grootswagers et al., 2017). The analyses were performed within the subjects and the subsequent analysis was performed at the group level. For our study, the representational dissimilarity matrices (RDMs) were constructed based on the dissimilarity patterns evoked by each stimulus pair (Kriegeskorte et al., 2008). With data from 200 images, the RDMs map out the dissimilarity patterns evoked by all stimulus pairs, i.e., for a total of $\binom{200}{2}$, i.e., 19,900 pairs.

For our decoding analyses, the voltages of all 64 EEG channels were used as features for each time point. A regularized (λ=0.01) linear discriminant classifier ($\lambda = 0.01$) ($\lambda = 0.01$) ($\lambda = 0.01$) was trained to distinguish between patterns evoked by different pairs of images. To assess the classification accuracy, a leave-one-sequence-out cross-validation procedure was used. Here, an image presentation sequence from each category was used as test data, while the classifier was trained on the remaining image presentation sequences. This resulted in (19,900 image condition pairs x 275 EEG time points) shaped EEG-RDM containing the mean classification accuracy scores for the image pairs in the left-out sequences for each subject. Finally, the RDMs were averaged across all image pairs to calculate the mean pairwise classification accuracy over time.

Layer Activation Patterns from DNNs

We chose foundational models in all modalities, that is, language, vision, and multimodal, comprising a subset of model architectures used in Choksi et al. (2022), since our hypothesis was to estimate the reliability of the same underlying hypothesis tested in Choksi et al. (2022) comparing model modalities and better explaining the representational dynamics of specific regions.

The multimodal architectures for our analysis included CLIP and VirTex. While CLIP is trained on contrastive learning (Radford et al., 2021), VirTex (Desai & Johnson, 2021) is trained on image captioning and constitutes a proper subset of architectures across various training paradigms as used in Choksi et al. (2022). For activity patterns, we extracted the latent representations from the attention pool layer from CLIP and the average pooling layer for Virtex from the visual backbone of either model.

Visual foundation model set comprises of models from vanilla ImageNet trained models like ResNet-50 (Julian et al., 2012) and BiT-M (Kolesnikov et al., n.d.), adversarially robust models (AR-L2, AR-L4, AR-L8) (Salman et al., 2020), and stylized ImageNet models, a model pre-trained on only stylized and original ImageNet images, a model trained on stylized and original ImageNet, and another model further fine-tuned on ImageNet post-training (Geirhos et al., 2018). Adversarially robust models are more human-like in their behavior, i.e., object detection performance is more resilient to OOD shifts due to adversarial noise in the dataset. Including stylized models, since the inductive biases introduced by training on regular ImageNet samples are more tuned toward texture specificities than other more prominent shape-reliant object features (Geirhos et al., 2018). For models with a visual foundation, the representations from the average pooling layer were used to create the RDMs.

Estimating Representational Dissimilarity Matrices

For EEG and fMRI, RDMs were created using different methodologies, replicating the paradigm followed by the

original authors in respective data modalities. The RDMs for the EEG data were created using the pairwise classification accuracy of the linear discriminant analysis model with a higher accuracy that indicates greater separability, translating into greater dissimilarity for the given pair of categories (Kaniuth & Hebart, 2022). For fMRI and DNN, response patterns for categories were compared using Pearson's correlation, with lower scores corresponding to higher dissimilarity between categories.

Comparing Representations across Model Modalities

Latent representations across model modalities demonstrate varying degrees of correspondence to brain response depending on the region of interest. Similar results have been reported in Conwell et al. (2024); Wang et al. (2023); Storrs et al. (2021); Oota et al. (2025), where depending on the training domain, viz. images, texts, or both, latent representations from a model better correlate to the elicited pattern as recorded in fMRI BOLD signals. For our analysis, we also constrained our comparisons across the four regions, viz. visual region, fusiform face region, hippocampus, and parahippocampal region. For computing model correlation with a particular brain region, we used mean Pearson's correlation coefficient across subjects between individual brain region RDMs and model RDMs. Similarly to Choksi et al. (2022), the correlation score for each model within a modality was used as independent samples for a modality.

Specification Curve Analysis

Alternate Specifications for fMRI

Our primary goal for both the fMRI specifications curve analysis was to evaluate analytical forks post-preprocessing. Our alternate specifications were tuned for either of the datasets depending on the respective analysis goals and data availability. Our analysis emphasized in particular the reliability of various normalization approaches and ROI mask estimation methods, as these steps have previously been linked to sources of noise in the estimation process for fMRI data processing pipelines(Ramírez, 2017; Murphy et al., 2009; Viswanathan et al., 2012; Friston et al., 2006; Duncan et al., 2009).

For the THINGS fMRI pipeline, the original pipeline implemented signal normalization on two scales: runwise and global. Although the fMRI literature presents various normalization techniques, the original analysis employed a percent signal change for run-wise normalization. In our comprehensive analysis, we explored three normalization methods at both the run-wise and global levels: percent signal change, z-standardization, and mean centering (demeaning). Additionally, we introduced a fourth option for global normalization—omitting this step completely, given that the signal had already undergone local normalization.

Furthermore, we explored alternative approaches for HRF selection and ROI mask estimation. For HRF selection in each region of interest, while the original pipeline was selected based on the mean R^2 between sessions, we developed an alternative method that selects HRF based on the function that produces the highest R^2 between sessions the most frequently. For ROI mask estimation, we created contrast maps by sampling exhaustive combinations of 2, 3, 4, and 5 runs from the six available runs, resulting in $\binom{6}{k}$ contrast maps for k=2,3,4,5. For each k, we generated two individual-level masks from these contrast maps, one through the union and another through the intersection of the maps. This approach yielded six alternative region masks in addition to the original mask, allowing us to evaluate the reliability of mask estimation methods for downstream RSA inference, particularly given the high degree of variability observed across sessions in active voxels and the limited sample size of only six runs. These methodological choices in normalization, HRF selection, and mask estimation resulted in 336 distinct analytical configurations including the original pipeline, the details of which are described in Table 1.

Our specification curve analysis for this dataset (Horikawa & Kamitani, 2017) focused exclusively on modifying the analysis choices during the response estimation phase, as we were unable to include variations in masking methods, as it was not possible to employ data-driven mask estimation for regions such as the hippocampus and parahippocampus. Therefore, our specification alternatives were confined to the post-preprocessing stages up through fMRI response estimation. We maintained consistency with our previous specification curve analysis for RDM consistency by implementing the same alternative configurations. The

Table 1 Alternative specifications for fMRI analysis

RTN	GTN	SR	BHRF	RU	MM
PSC	RCFM	ZSR	H	6R	I
Z-S	RPSC	NZSR	HM	3R	U
CFM	RZ-S			4R	
	OF			5R	

RTN: *PSC* % signal change, *Z-S* Z-standardization, *CFM* mean centering; GTN: *RCFM* Mean centering, *RPSC* % Signal change, *RZ-S* Z-Standardization, *OF* None; SR: *ZSR* Z-scoring, *NZSR* None; BHRF: *H* Using mean, *HM* Using mode; RU: *R* runs; MM: *I* Intersection, *U* Union. 1st row corresponds to original specification

analysis choices included run-wise signal scaling (percent signal change, data demeaning, z standardization, and no normalization as a new addition), global signal normalization (percent signal change, data demeaning, z standardization, and the option to skip global rescaling given prior run-wise scaling), and HRF selection of best fit based on explained variance (R^2) using either mean R^2 across runs or frequency-based selection (choosing HRF that performed best most often across runs). Based on the results of our previous analysis, we removed a particular set of combinations for normalization, viz. performing a global percent signal change transformation only when the run-wise normalization step was skipped. These methodological choices yielded 26 distinct pipeline configurations ($4 * 4 * 2 = 32 - (3 * 2) = 26$). Unlike our previous specifications curve analysis that examined the robustness of the RDM, we did not incorporate the original pipeline specification in this analysis due to the availability of the pipeline recipe. In addition to all these, we included an alternate specification for region masks, i.e., using the region masks directly, instead of employing the top-30 most responsive voxel selection. Thus, our analysis included 52 alternate specifications in total.

Alternate Specifications for EEG

We devised our specification sets by altering the decisions involved in pairwise classification accuracy for EEG for the whole brain and region-specific electrode maps. We devised configurations from channel selection to estimate cross-validation folds for accuracy for EEG analysis. The region electrode combinations included a diverse set with each of the four regions, that is, frontal, central, temporal, and parietal-occipital, constituting forks of the specifications tree, and the whole brain electrode as a separate fork. The cross-validation fold for RSA in EEG identifies the sensitivity of the LDA classifier to varying exemplars for a particular pair of conditions, identifying the classifier's reliability in discriminating the characteristics latent in the EEG signal. We used an exhaustive set of folds for cross-validation. Here, k sequences, ranging from 1 to 10, were reserved as test data while training the classifier on the remaining (12-k) image presentation sequences for each category. To further assess the robustness of our findings post-decoding, we focused our SCA on RDMs generated at four different time points (150 ms, 200 ms, 250 ms, and 300 ms post-stimulus) and also averaged the RDMs across the entire post-stimulus time window (0 to 996 ms). This gave us five different sets of RDM specification for performing our SCA in the post-decoding stage. In total, 10 * 5 * 5 = 250 specifications, including the original specification that used whole brain electrodes along with leave-one-sequence cross-validation, and the RDM generated by five different temporal profiles were used as alternate specifications.

Results

Our first objective was to test how pipeline configurations influence the significance of estimated RDM across neuroimaging modalities. To this end, we present two sets of results: one for task-based fMRI and one for EEG data. And, for fMRI data, we further demonstrated how model hypotheses tested using RSA are influenced by the analysis choices using a subset of our original specification. Finally, we also show a novel data-driven approach for identifying robust configurations for neuroimaging analysis and how the same results in reliable pipeline configurations for analysis pipelines with high degrees of freedom, such as fMRI-based RSA.

Robustness of fMRI-Based RSA to Alternate Specifications

We evaluated the significance of the RDMs resulting from each pipeline configuration using a two-sided significance test for rank correlation, Kendall's τ_A. Kendall's τ_A was estimated by comparing the RDMs resulting from each pipeline configuration with the RDM resulting from the original configuration. Using the 19 shuffled samples, we estimate the p-value of the RDMs from each pipeline configuration. Finally, for an RDM from an alternate specification to be significant, similarity to the RDM from the original specification (here Kendall's τ_A) for the original data in comparison to the shuffled samples must cross the threshold corresponding to the significance level(α).

Our results (see Fig. 1) showed significant RDM estimates at α=0.025 for 126 of the total of 335 alternative specifications, with the original specification trivially true.

The specification curve for the analysis is illustrated, showing that in roughly 65% of the alternate specifications, the mean response patterns resulting from the alternate specifications used to create the RDMs are not significantly different from those estimated from the null samples. This result highlights that the sparse sampling (Turner et al., 2018) used in most RSA studies is prone to introduction of bias in the pipeline due to the compositional analytic interaction with task-based fMRI with low SNR.

We also found that the original specification is quite fragile due to changes in the analysis choices. With the change in one decision point for any specification choices, we have analyzed (six in total), resulting in almost half of pipeline configurations producing nonsignificant RDMs. Of the total possible 11 analysis choices, changes to only six (see Fig. 7 for a split of each analysis choice), keeping every other configuration unchanged, produced correlations significantly different from the null distribution.

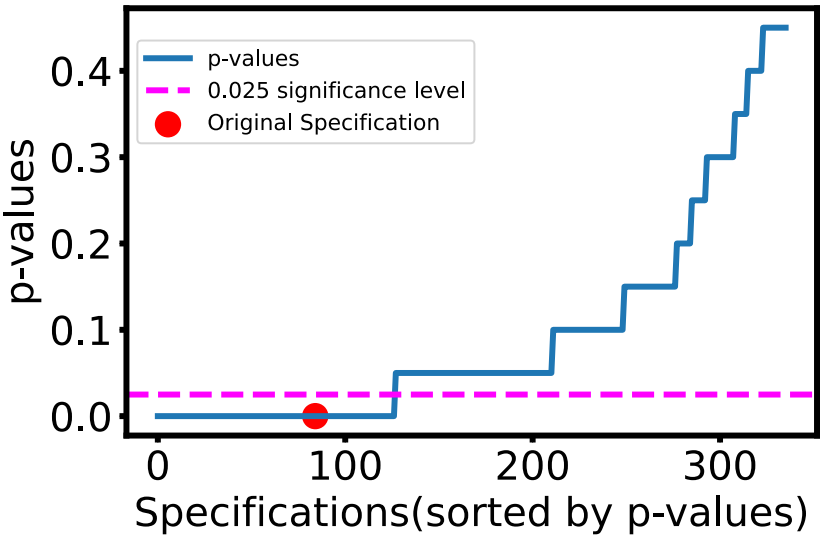

Fig. 1 Specifications curve for fMRI

Robustness of EEG-Based RSA to Alternate Specifications

For EEG, the significance of the RDMs resulting from pipeline configurations was also evaluated using a two-sided significance test for rank correlation, Kendall's τ_A. Kendall's τ_A was used to statistically evaluate the differences between the representational dissimilarity matrices (RDM) of our original, unshuffled EEG data and those derived from the shuffled samples in every pipeline configuration. We created a baseline for comparison by randomly shuffling the stimulus labels in the time series data for each stimulus presentation, and we repeated this process 10 times to construct a null distribution. For each shuffle, we calculated the subject-wise representational dissimilarity matrices (RDMs) and estimated their mean time-varying decoding accuracy. We then calculated the p values by comparing the representational dissimilarity matrices (RDMs) of the original data with each pair of k values of the shuffled samples using Kendall's τ_A.

The analysis showed significant results [$p < 0.025$] (see Fig. 2) for all 250 specifications, indicating the high robustness of RSA based on EEG. This robustness was observed in the configuration space of 10 leave-k-sequence-out cross-validation folds (k = 1:10), the five temporal RDM profiles (150 ms, 200 ms, 250 ms, 300 ms, and the average of 0–996 ms), and both the complete 64 channel montage and the region-based electrode groupings (frontal, central, temporal, and parietal occipital).

The decoding model consistently produced significant results for all specifications, indicating that the model is stable and reliable across various electrode, cross-validation, and temporal configurations. This suggests that the model successfully captures the underlying neural patterns that

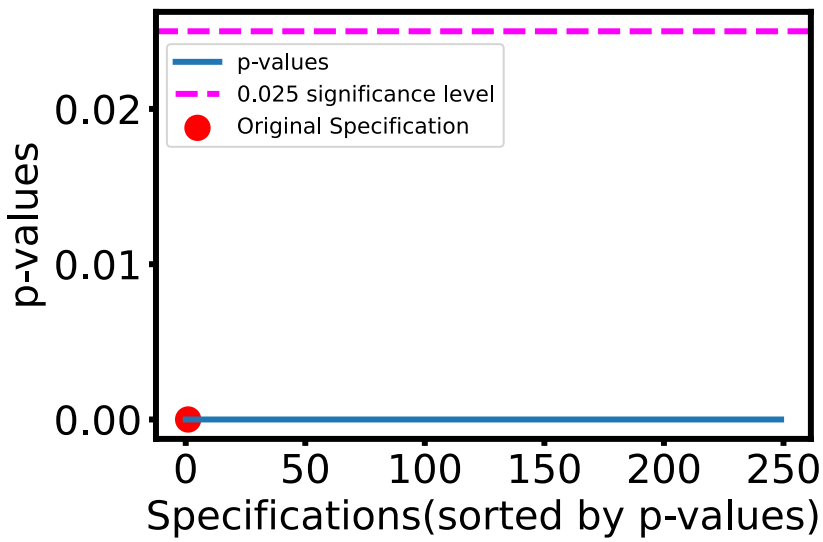

Fig. 2 Specifications curve for EEG

correspond to the stimuli presented. Unlike fMRI, which showed robustness for only 126 of 335 specifications, the alternative specification curve analysis for EEG indicates a high degree of robustness across all choices in the pipeline. Consistent, significant results across cross-validation configurations, electrode groupings, and temporal profiles indicate the generalizability of the model, which shows that it reliably decodes EEG data without overfitting. These findings highlight the robustness of EEG decoding pipelines for analyzing the dynamics of neural representation (Table 2).

Reliability of Model-Based Hypothesis Evaluation Using RSA

The primary analysis of Choksi et al. (2022) investigated whether any particular class of models is more similar to another with the individual brain regions, namely the fusiform area, the visual region, the parahippocampal area, and the hippocampus. Specifically, the hypothesis of interest was involving the hippocampal region and the performance of multimodal model class in comparison to the other two for that particular region—models like CLIP and VirTex with both language and vision backbones are better in predicting the response geometry than unimodal vision models like ResNet-50 and language models like BERT and GPT-2. The final statistical *t*-test was performed on normalized RSA scores, which were calculated using Pearson's correlation coefficient between the subject RDM and the model RDM. The score was divided with the subject-specific noise lower bound for the particular region before averaging among the participants to produce the final score for each model. Although the noise ceiling was considerably high for the other three regions, viz., visual region, fusiform, and parahippocampal area in the range of 0.2−0.6, it was much lower for the hippocampus, around 0.1 and 0.15 even with the 30 top voxel selection paradigm, which improved the noise ceiling bounds over the original region-specific voxel set. Thus, this set of tests encompassing a wide range of normalized RSA scores with varying ranges of noise ceilings provides a basis for reevaluating the hypothesis of the RSA-based model in regions with varying degrees of signal-to-noise ratio (SNR).

We have presented the results of the original analysis in Table 3. Analysis in brain regions revealed consistent advantages for multimodal architectures compared to visual and language models, with visual models demonstrating stronger effects than language models. The fusiform area exhibited notable positive effects for both multimodal and visual models when compared against language models. In the hippocampal region, a clear hierarchical pattern emerged—multimodal models showed the strongest effects, followed by visual models, and then language models, with significant differences between all three classes. In particular, only multimodal models reached the threshold for the hippocampus noise ceiling, a feat not achieved in other brain regions. This comprehensive analysis was subsequently validated by replication using a selective subset of our predefined analytical specifications to confirm the observed modality-specific advantages in relation to brain region RDMs.

We re-tested each of the region-specific statistical tests using our set of alternate specifications to estimate the difference in modality-specific model advantages when compared with brain region RDMs.

Analysis of the fusiform region revealed differential effects across model modalities in the initial investigation. Although a positive systematic trend emerged from multimodal to unimodal models, statistical significance was limited to comparisons with language models, where both visual and vision language models demonstrated higher RSA scores. However, our analysis of the specification curve (Fig. 3) revealed no robust statistically significant effects in 52 pipeline configurations.

Table 2 Alternative specifications for the EEG analysis pipeline

Electrode selection	Cross-validation strategy	RDM time window
Whole brain	Leave-1-out	200ms
Frontal	Leave-2-out	150 ms
Central	Leave-3-out	250 ms
Temporal	Leave-4-out	300 ms
Parietal-Occipital	Leave-5-Out	Averaged (0–996ms)
	Leave-6-out	
	Leave-7-out	
	Leave-8-out	
	Leave-9-out	
	Leave-10-out	

Each unique pipeline is formed by selecting one option from each of the three parameter columns. 1st row corresponds to original specification

Table 3 Comparison of effects across regions from (insert ref)

Region	MxV	MxL	VxL
Visual region	Positive	Positive	Positive
Fusiform area	Positive	Positive(*)	Positive(*)
Hippocampus	Positive(*)	Positive(*)	Positive(*)
Parahippocampus	Positive	Positive	Positive

* indicates significant differences between the classes ($p < 0.05$). *MxV*, multimodal models vs vision models; *MxL*, multimodal models vs language models; *VxL*, vision models vs language models. Positive effect indicates that the effect was higher for the first model class on the column label, i.e., if MxV is the header, then the positive effect indicates that the multimodal model class, on average, has a more positive similarity score with brain RDMs than vision models

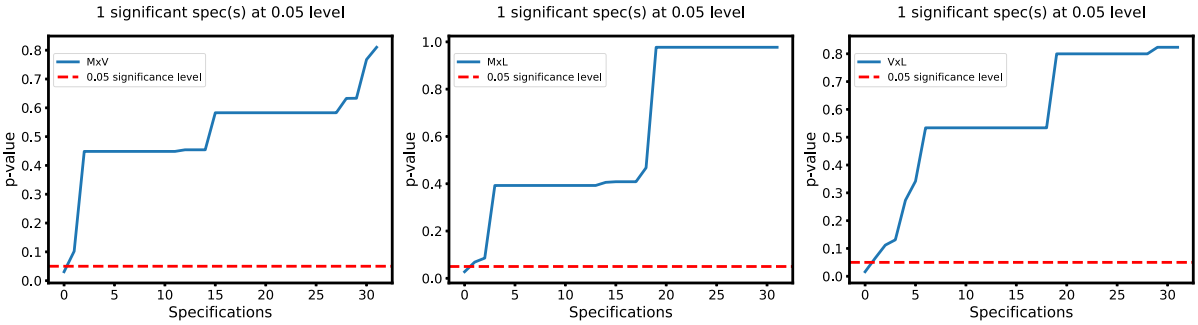

Fig. 3 Specifications curve for model class comparisons for fusiform area RDMs. Positive effect indicates that the effect was higher for the first model class on the column label, i.e., if MxV is the header, then the positive effect indicates that the multimodal model class, on average, has a more positive similarity score with brain RDMs than vision models

The original analysis (Table 3) showed decreasing but nonsignificant similarity between visual region RDM and model classes, from multimodal to visual to language models. Our specification curve analysis revealed 13 of 52 pipeline configurations that demonstrated significant effects ($p < 0.05$) following this trend in all pairwise comparisons, as shown in Fig. 4. Notably, multi-modal vision-language models showed a significantly stronger correlation with visual region representations compared to unimodal vision models. In essence, about 40% of specifications revealed consistent significant undetected effects in the original analysis.

In the original analysis, the parahippocampal region revealed a pattern of decreasing but nonsignificant similarity across model classes, with multi-modal models showing the strongest correlation, followed by visual models and then language models. However, our specifications curve analysis painted a notably different picture, demonstrating the variability one might observe by employing alternate pipeline configurations, as shown in Fig. 5. Firstly, one set of specifications (13 in total) demonstrated a significant effect for multi-modal models compared to visual models. The language models set also produced a net positive effect difference to visual models for that set of specifications but did not cross the significance threshold. Another group of specifications (13 in total) displayed an effect entirely in contrast to this—with visual models producing significant similarity effect difference compared to language and multi-modal models in 10 configurations. The comparison of multi-modal to language models was also significant for those specifications, with the latter performing worse. In addition, a similarly significant effect difference in favor of visual models was observed from three more specifications. Finally, another group of pipeline configurations (13 in total) produced only a significant effect difference for language models with visual models in favor of the former, with the multimodal model effect confined between the effects from two other model modalities with no significant difference with either.

Choksi et al. (2022) analysis revealed significant effects in all comparisons of the hippocampus model classes, with multimodal models showing the highest similarity, followed by visual and language models. Both multimodal models reached the neural RDM noise ceiling, albeit at notably low levels. However, our specifications curve analysis produced

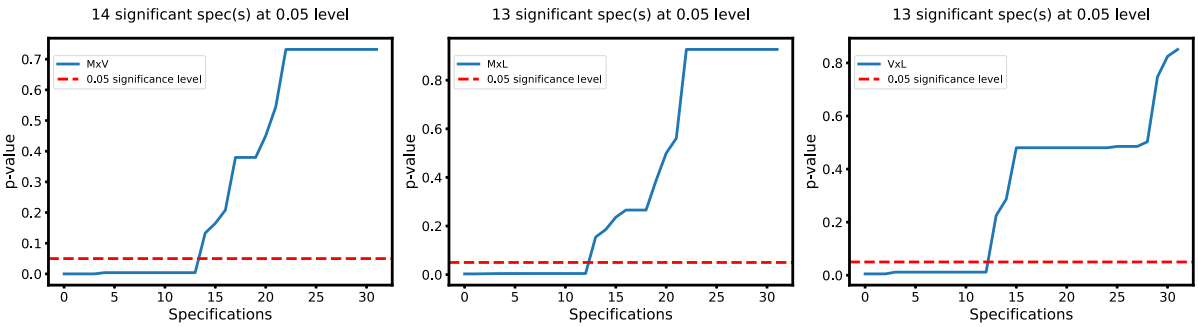

Fig. 4 Specifications curve for model class comparisons for visual region RDM. Positive effect indicates that the effect was higher for the first model class on the column label, i.e., if MxV is the header, then the positive effect indicates that the multimodal model class, on average, has a more positive similarity score with brain RDMs than vision models

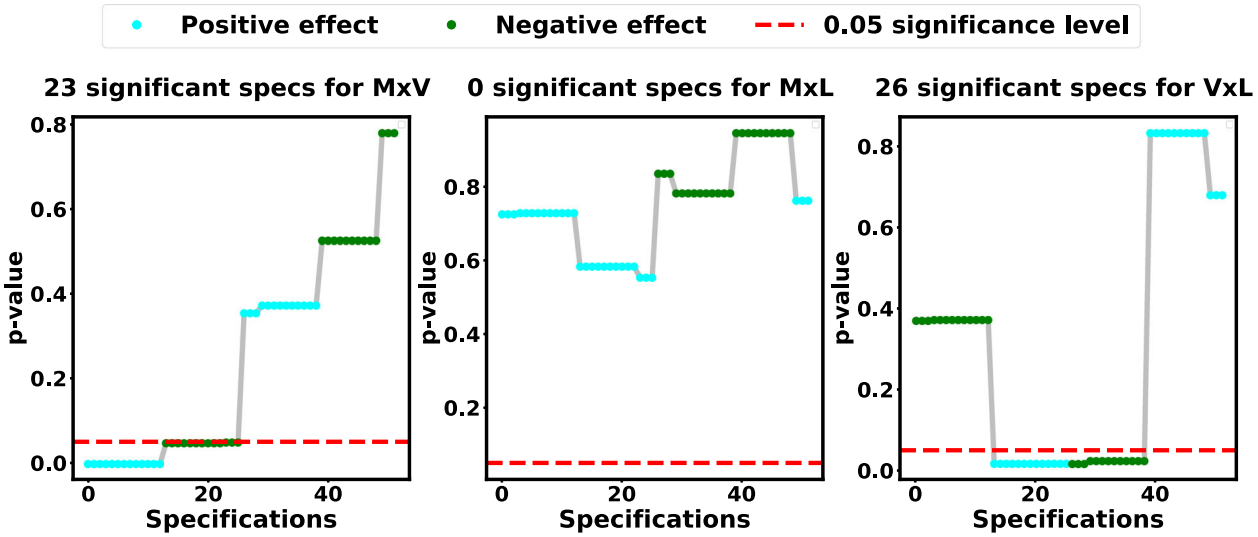

Fig. 5 Specifications curve for model class comparisons for parahippocampal area RDM. Positive effect indicates that the effect was higher for the first model class on the column label, i.e., if MxV is the header, then the positive effect indicates that the multimodal model class, on average, has a more positive similarity score with brain RDMs than vision models

results that were incongruent with the original results, as shown in Fig. 6. Although the directionality of the effect was consistent for most specifications, none was statistically significant. Moreover, half of the pipelines produced a significant effect difference for visual models compared to multi-modal models.

This cacophony of results seen across theoretically consistent alternative specifications presents clear evidence for caution in interpreting results of model-based hypothesis tests of intrinsically unreliable primary data.

Finding Robust fMRI Specifications

The set of alternative specifications defined in SCA is obtained by a branching process in the garden of analysis forks, which eventually yields specifications that are

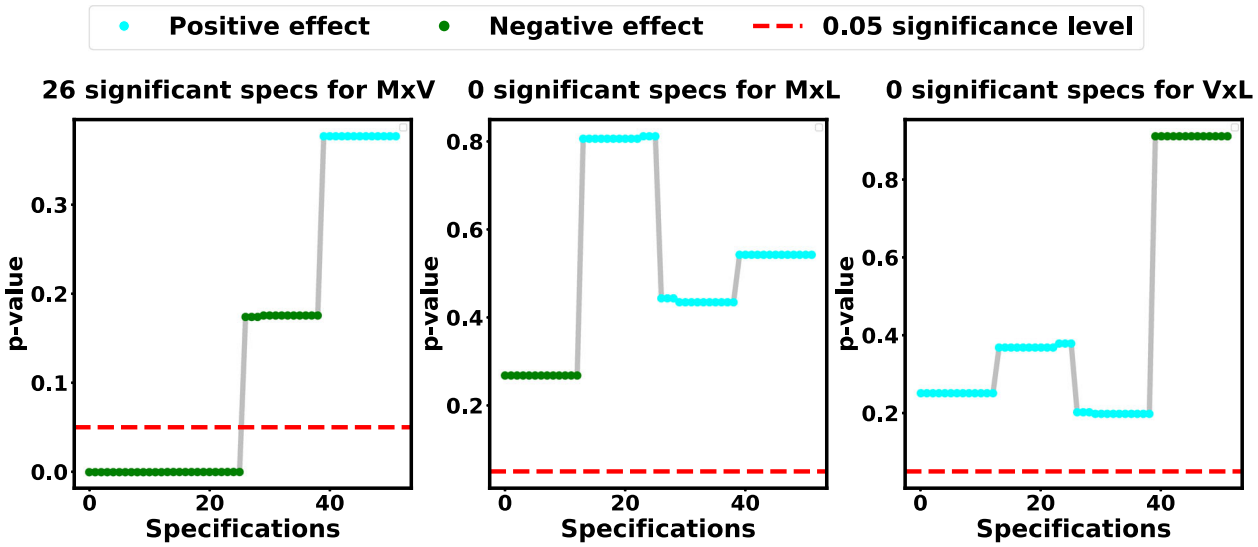

Fig. 6 Specifications curve for model class comparisons for hippocampal area RDM. Positive effect indicates that the effect was higher for the first model class on the column label, i.e., if MxV is the header, then the positive effect indicates that the multimodal model class, on average, has a more positive similarity score with brain RDMs than vision models

significantly or non-significantly different from the null distribution. Therefore, it is possible to model the mapping of specifications to the binary prospect of being significant or not by fitting a decision tree classifier (Quinlan, 1983) using the specification choices as features and significance as the binary target variable. Examination of the structure of the learned tree reveals the importance of each parameter specification based on frequency. The path of the decision tree leading to leaf nodes with the highest number of significant and nonsignificant specifications is tabulated in Tables 4 and 5, respectively. Evidently, using five runs seems to be the most predictive of significance for specifications as shown in Table 4, although even here, selecting some choices in other analysis forks can lead to clusters of nonsignificant specifications, as shown in Table 5.

Based on the decision tree's structure, we defined the most robust specification as the one that leads to the least number of non-significant specifications if one analysis choice is varied. For the set of alternative specifications we used, the specification changes for the most robust specification (PSC-OF-ZSR-H-5R-I) (see Fig. 7). For this specification, changes in either HRF estimation strategy, z-standardization of residuals, or normalization methods do not affect the significance of the correlations. Only using union instead of intersection for ROI definition results in a nonsignificant combination.

Compared with the original specification, this does not use any form of global normalization. Thus, by simplifying the original specification in this way, one can arrive at a specification that is substantially more robust to further alternative specifications, at least with reference to the set of specifications and dataset we have used.

Discussion

Neuroscience is replete with studies based on fMRI using stimulus-yoked designs to make claims about the location of various cognitive phenomena (Weisberg et al., 2008). Methodological concerns about Type-1 and Type-2 errors have been prominent in non-invasive neuroscience research (Pashler & Wagenmakers, 2012; Barch & Yarkoni, 2013). General concerns about false positives in a highly flexible data analysis pipeline have been widely accepted (Kelly Jr & Hoptman, 2022; Gelman & Loken, 2014). However, principled approaches to solving the problem have been hard to come by; current practice has primarily emphasized the use of preregistration and openness to publication of null results or failed replications (Nelson et al., 2018). However, recent empirical evidence from across disciplines shows that *p-hacking*, i.e., altering pipelines post-hoc to meet analysis goals, does not actually reduce unless preregistration is accompanied by predefined analysis plans (PAP) (Brodeur et al., 2024).

Table 4 Specification choices producing the largest number of significant correlations

RTN	GTN	SR	BHRF	RU	MM	#
Any	OF	Any	Any	3R	U	12
Any	RCFM	Any	Any	5R	I	12
Any	RZ-S	Any	Any	5R	I	12
Any	OF	Any	Any	5R	I	12
Any	PSC	Any	H	5R	U	6

Table 5 Specification choices producing the largest number of non-significant correlations

RTN	GTN	SR	BHRF	RU	MM	#
Any	Any	Any	Any	3R	I	48
CFN	Any	Any	HM	4R	Any	16
PSC	Any	Any	HM	4R	Any	16
Any	OF	Any	Any	5R	U	12
Any	PSC	Any	Any	5R	I	12

In this paper, we argue that for the same neural computational model assumptions, there may exist multiple combinations of valid analysis choices for encoding and decoding model estimation pipelines, all consistent with a declared high-level PAP. For example, multivariate encoding and decoding analyses in fMRI and EEG, respectively, work under a similar hypothesis of the underlying feature space to transform the representations of features from the underlying feature space to the activity space, forming the basis for the development of underlying encoding and decoding models (Naselaris et al., 2011). However, pipeline specification combinations span from these single underlying hypotheses to a multidimensional parameter space for both modalities, and more so for fMRI pipelines. Therefore, we suggest that evaluating the consistency of the results obtained in alternative specifications is necessary for inferences resulting from processing pipelines with many degrees of freedom, a characteristic of non-invasive neuroscience research. In this paper, we showed how to do this for one particular analysis, RSA.

In light of the large garden of forking paths that is known to exist in the processing of fMRI data for stimulus-linked experiments, we adapted specification curve analysis for representation similarity analyses based on fMRI data (Simonsohn et al., 2020), focusing on perturbing the analytical stages, which are both critical and epistemically fragile. We showed that, compared to a baseline EEG-based representation similarity analysis pipeline, the fMRI pipeline shows significant variability in the results of multiple theoretically reasonable alternative specifications. In particular, only a third of the set of alternative specifications achieved statistical significance in our analysis. In contrast, RSA using EEG

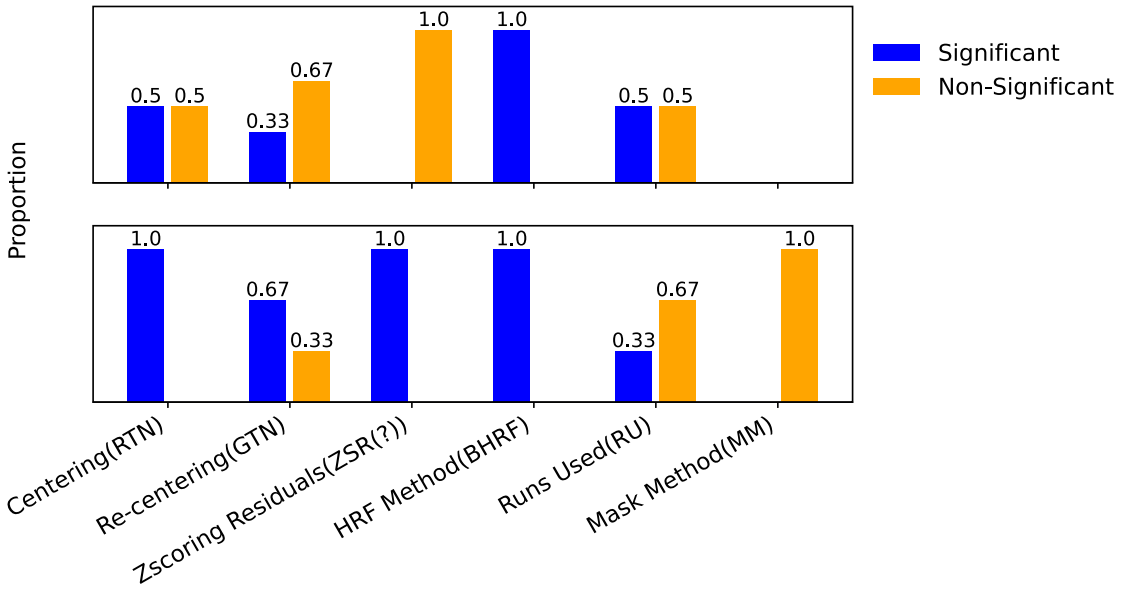

Fig. 7 Proportion of analysis pipelines producing significant RDMs. Top panel shows how shifting each analysis choice results in significance (for the original pipeline configuration). Bottom panel shows how shifting each analysis choice results in significance (for our identified most robust pipeline configuration)

as primary data proved to be substantially robust to alternative specifications. We also found that model-based analysis using fMRI data inherits and even amplifies the frailty of the primary data to alternative specifications, such that almost no conclusions from the original analysis can be reliably supported by a specifications curve analysis. We finally showed, using a decision tree-based approach, how a small change in the original specification could make it more robust.

However, we note that even the most robust specification identified in our exercise would still not reduce the overall fragility of the fMRI analysis pipeline. For example, switching the mask estimation method from I to U in the robust specification discovered in our analysis yields a large number of nonsignificant alternatives, as evident by the presence of this modified specification in Table 2. Moreover, we could not identify any theoretically sensible pattern to characterize specifications failing to pass the significance test based on SCA. Thus, we conclude that, in addition to its known unreliability (Elliott et al., 2020), single-test fMRI data is also fragile to the analysis choices in its information processing pipeline. Therefore, caution is warranted in interpreting representational similarity results using fMRI data.

In our analysis, given the combinatorially large space of possible configurations emerging from the alternate choices at various stages of the analytical pipeline in task-based fMRI, the specification set we selected necessarily represents a modest subset of the set of all theoretically reasonable analysis specifications—from noise normalization to choice of dissimilarity measures in RSA—a multitude of methods can apply towards the same analysis goals.

In particular, in our SCA analysis for fMRI, we constructed the set of requirements by focusing on analytical possibilities after pre-processing in the respective pipelines. The specification set constructed for both analyses in fMRI included only one set of noise regressors to remove non-experimental noise artifacts from raw fMRI time series.

Other possible combinations of anatomical components, drift regressors, ICA regressors, and other features that contribute to physiological and motion artifacts could also be used to construct the final nuisance regressors. For example, using a different degree of polynomial for modeling drift artifacts, or using anatomical component correction (aCompCor), explaining half of the variance among the principal components, or even using alternate thresholding criteria for filtering the ICA regressors, are all theoretically salient approaches toward the same goal, discriminating noise from underlying neural patterns (Behzadi et al., 2007; Charest et al., 2018). Moreover, other methods, such as multivariate and univariate noise normalization methods, are also prevalent and often used interchangeably in processing pipelines (Walther et al., 2016; Ritchie et al., 2021).

Similarly, for another core component of the pipeline, namely, statistical modeling, there exist multiple alternatives in addition to the ridge regression method employed for the fast event-related design, such as least squares separate (LSS) (Mumford et al., 2012; Turner et al., 2012), inverse transformed encoding model (ITEM) (Soch et al., 2020), and GLM variants with alternative regularization techniques, such as alternative regularization techniques such as LASSO (Gaudes et al., 2011).

HRF estimation methods also demonstrate similar flexibility in their choice with effervescence of methods that coexist in the fMRI literature (Lindquist et al., 2009). In our specification set, we used a library of 20 HRFs, derived from FIR models on a large-scale dataset to model individual differences (Prince et al., 2022), and later fine-tuned them in regions or voxels for analyses. However, pipelines often leverage FIR models, sFIR models, alternative basis sets (like spectral, spline, or gamma functions), and even canonical HRF and its derivatives for HRF estimation. In addition to the plethora of options emerging from combinatorial juxtaposition of the alternatives available across the various analytical stages, there exist end-to-end data-driven modeling techniques like GLMDenoise (Charest et al., 2018), encapsulating the entire pipeline from denoising to response estimation. Further downstream in the analysis pipeline, voxel selection also presents a handful of alternatives in addition to data-driven region masks from category localizers—using various parcellations for ROIs from pre-existing atlases, searchlight analyses (Kriegeskorte et al., 2006), or other data-driven parcellation schemes (Degryse et al., 2017; Parmar et al., 2022). Finally, for RSA, the dissimilarity measure itself presents another potential source of bias (Walther et al., 2016), adding another degree of potential flexibility in fMRI response estimation.

Fundamentally, in any specification curve analysis (SCA), analysts make their own choice of which set of specifications to include and which to exclude in the analysis (Simonsohn et al., 2020). This choice can itself be questioned on grounds of arbitrariness. However, it is important to note that the value of the SCA analysis is not based on the use of a comprehensive set of specifications but rather on the ability to identify how the results fluctuate as a function of changes in specifications (Simonsohn et al., 2020).

Including one or more of the alternatives listed above in the set of specifications may individually have made some difference to the trajectory of the specification curve, but it is highly unlikely that including them would have changed its overall shape for either measurement modality.

We conclude with the observation that SCA could and should be applied without loss of generality to nearly all event-locked fMRI data analysis protocols, with suitable modifications to the set of alternative specifications we used in our work.

Author Contributions NS conceptualized the study, SS and GB conducted analyses and tested models, and SS, GB, and NS wrote the paper.

Funding Not applicable

Data Availability No datasets were generated or analyzed during the current study.

Materials Availability Not applicable.

Code Availability Source code for the methods described in this paper can be accessed at our code repository.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication All authors approve the manuscript for publication.

Conflict of Interest The authors declare no competing interests.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., et al. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience, 25*(1), 116–126.

Badillo, S., Vincent, T., & Ciuciu, P. (2013). Group-level impacts of within-and between-subject hemodynamic variability in fMRI. *Neuroimage, 82*, 433–448.

Barch, D. M., & Yarkoni, T. (2013). *Introduction to the special issue on reliability and replication in cognitive and affective neuroscience research* (Vol. 13). Springer.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage, 37*(1), 90–101.

Brodeur, A., Cook, N. M., Hartley, J. S., & Heyes, A. (2024). Do preregistration and preanalysis plans reduce p-hacking and publication bias? Evidence from 15,992 test statistics and suggestions for improvement. *Journal of Political Economy Microeconomics, 2*(3), 527–561.

Carlson, T., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision, 13*(10), 1–1.

Charest, I., Kriegeskorte, N., & Kay, K. N. (2018). GLMdenoise improves multivariate pattern analysis of fMRI data. *Neuroimage, 183*, 606–616.

Choksi, B., Mozafari, M., Vanrullen, R., & Reddy, L. (2022). Multimodal neural networks better explain multivoxel patterns in the hippocampus. *Neural Networks, 154*, 538–542.

Clerc, M., Bougrain, L., & Lotte, F. (2016). *Brain-computer interfaces 1: Methods and perspectives*. John Wiley & Sons.

Cohen, M. X. (2017). Where does EEG come from and what does it mean? *Trends in Neurosciences, 40*(4), 208–218.

Cole, S., & Voytek, B. (2019). Cycle-by-cycle analysis of neural oscillations. *Journal of Neurophysiology, 122*(2), 849–861.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, *15*(1), 9383.

Degryse, J., Seurinck, R., Durnez, J., Gonzalez-Castillo, J., Bandettini, P. A., & Moerkerke, B. (2017). Introducing alternative-based thresholding for defining functional regions of interest in fMRI. *Frontiers in Neuroscience, 11*, 222.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9–21.

Desai, K., & Johnson, J. (2021). Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11162–11173).

Duncan, K. J., Pattamadilok, C., Knierim, I., & Devlin, J. T. (2009). Consistency and variability in functional localisers. *Neuroimage, 46*(4), 1018–1026.

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S. et al. (2020). What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychological Science, 31*(7), 792–806.

Fischl, B. (2012). Freesurfer. *Neuroimage, 62*(2), 774–781.

Friston, K. J., Rotshtein, P., Geng, J. J., Sterzer, P., & Henson, R. N. (2006). A critique of functional localisers. *Neuroimage, 30*(4), 1077–1087.

Gaudes, C. C., Petridou, N., Dryden, I. L., Bai, L., Francis, S. T., & Gowland, P. A. (2011). Detection and characterization of single-trial fMRI bold responses: Paradigm free mapping. *Human Brain Mapping, 32*(9), 1400–1418.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint* arXiv:1811.12231.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460–465.

Gramfort, A., Strohmeier, D., Haueisen, J., Hämäläinen, M. S., & Kowalski, M. (2013). Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations. *Neuroimage, 70*, 410–422.

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience, 29*(4), 677–697.

Grootswagers, T., Zhou, I., Robinson, A. K., Hebart, M. N., & Carlson, T. A. (2022). Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data, 9*(1), 3.

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience, 7*(7), 523–534.

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A. et al. (2023). THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife, 12*, e82580.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS ONE, 14*(10), e0223792.

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications, 8*(1), 15037.

Inc., T. M. (2020). *Matlab version: 9.9.0 (r2020b)*. Natick, Massachusetts, United States: The MathWorks Inc. https://www.mathworks.com

Julian, J. B., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage, 60*(4), 2357–2364.

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

Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F., & Wandell, B. A. (2013). GLMdenoise: A fast, automated technique for denoising task-based fMRI data. *Frontiers in Neuroscience, 7*, 247.

Kelly Jr, R. E., & Hoptman, M. J. (2022). *Replicability in brain imaging* (Vol. 12) (No. 3). MDPI.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., & Houlsby, N. (n.d.). Big transfer (bit): General visual representation learning. arxiv 2020. *arXiv preprint* arXiv:1912.11370.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences, 103*(10), 3863–3868.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 4*.

Lindquist, M. A., Loh, J. M., Atlas, L. Y., & Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling. *Neuroimage, 45*(1), S187–S198.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving bold activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage, 59*(3), 2636–2643.

Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., & Bandettini, P. A. (2009). The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? *Neuroimage, 44*(3), 893–905.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage, 56*(2), 400–410.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*(1), 511–534.

Oota, S. R., Pahwa, K., Marreddy, M., Singh, M., Gupta, M., & Raju, B. S. (2025). Multi-modal brain encoding models for multi-modal stimuli. *arXiv preprint* arXiv:2505.20027.

Parmar, H., Nutter, B., Long, R., Antani, S., & Mitra, S. (2022). Functional parcellation of fMRI data using multistage k-means clustering. *arXiv preprint* arXiv:2202.11206.

Pashler, H., & Wagenmakers, E.-J. (2012). Replicability in psychological science: A crisis of confidence?[special section]. *Perspectives on Psychological Science, 7*(6), 10–1177.

Pedregosa, F., Eickenberg, M., Ciuciu, P., Thirion, B., & Gramfort, A. (2015). Data-driven HRF estimation for encoding and decoding models. *NeuroImage, 104*, 209–220.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *Neuroimage, 45*(1), S199–S209.

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022). Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *Elife, 11*, e77599.

Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess end games. In *Machine learning* (pp. 463–482). Elsevier.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Ramírez, F.M. (2017). Representational confusion: The plausible consequence of demeaning your data. *bioRxiv,* 195271.

Ritchie, J. B., Masson, H. L., Bracci, S., & de Beeck, H. P. O. (2021). The unreliable influence of multivariate noise normalization on the reliability of neural dissimilarity. *NeuroImage, 245*, 118686.

Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., & Madry, A. (2020). Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems, 33*, 3533–3545.

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour, 4*(11), 1208–1214.

Soch, J., Allefeld, C., & Haynes, J.-D. (2020). Inverse transformed encoding models-A solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding. *NeuroImage, 209*, 116449.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience*, *33*(10), 2044–2064.

Turner, B. O., Mumford, J. A., Poldrack, R. A., & Ashby, F. G. (2012). Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. *NeuroImage, 62*(3), 1429–1438.

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology, 1*(1), 62.

Viswanathan, S., Cieslak, M., & Grafton, S. T. (2012). On the geometric structure of fMRI searchlight-based information maps. *arXiv preprint* arXiv:1210.6317.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage, 137*, 188–200.

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, *5*(12), 1415–1426.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*(3), 470–477.