Imaging Neuroscience
an open access journal

Research Article

Nonlinear KCCA in fMRI activation analysis: Self-supervised optimization and robust back-reconstruction

Chendi Han[a], Zhengshi Yang[a], Xiaowei Zhuang[a], Dietmar Cordes[a,b]

[a]Cleveland Clinic Lou Ruvo Center for Brain Health, Las Vegas, NV, United States
[b]Department of Psychology and Neuroscience, University of Colorado, Boulder, CO, United States

Corresponding Author: Dietmar Cordes (cordesd@ccf.org)

ABSTRACT

Recent studies have extended nonlinear kernels to Kernel Canonical Correlation Analysis (KCCA), enabling more flexible modeling of complex relationships globally. Building on these developments, we propose three key enhancements to nonlinear KCCA. First, inspired by self-supervised learning in machine learning research, we refine the parameter optimization process by adopting a subject-wise criterion designed to mitigate overfitting. Second, we introduce an improved back-reconstruction (inverse mapping) method that achieves higher accuracy and robustness than existing voxel-importance estimation methods. Third, we further investigate the kernel selection strategy based on convergence behavior, and validate its effectiveness through activation accuracy, data augmentation robustness, and eigendecomposition. The proposed framework is evaluated on both simulated and task-based fMRI datasets, with results demonstrating consistent improvements across multiple performance metrics.

Keywords: data analysis, fMRI, task fMRI, activation, nonlinear kernel, CCA, KCCA

1. INTRODUCTION

The task-fMRI activation detection problem aims to identify activation regions associated with a particular task design. From a methodological perspective, there are two primary ways to categorize existing methods. First, based on the input dimensionality, the methods can be classified as either local or global. Local methods, such as the General Linear Model (GLM) combined with a Gaussian Smoothing (GS) filter (Friman et al., 2003) or local Canonical Correlation Analysis (CCA) (Cordes et al., 2012; Friman et al., 2001; Yang et al., 2025; Zhuang et al., 2020), are computed in each local neighborhood around a voxel of interest. Global methods, such as Kernel Canonical Correlation Analysis (KCCA) and other techniques, detect activation in the entire brain in a single step (Han et al., 2025; Hardoon et al., 2004; Yang et al., 2018). Another classification considers whether the approach is linear or nonlinear. Linear methods, including GLM with GS (Friman et al., 2003), local CCA (Friman et al., 2001), and the linear KCCA (Hardoon et al., 2004; Yang et al., 2018), only capture linear relationships between the fMRI signal of the neural response and the hypothetical Hemodynamic Response Function (HRF), which is insufficient for HRF variability across different subjects or different voxel regions within a subject (Glover, 1999; Lindquist et al., 2009). Nonlinear approaches, such as local CCA with constraints (Cordes et al., 2012; Zhuang et al., 2020), deep CCA (Yang et al., 2025), or nonlinear KCCA (Han et al., 2025), can accommodate more general types of relationships.

One major challenge for global methods is the high dimensionality of unknowns relative to the available

Received: 3 September 2025 Revision: 21 April 2026 Accepted: 22 April 2026 Available Online: 28 April 2026

The MIT Press
© 2026 The Authors. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

supervisory signals. Traditional task-fMRI data have a spatial dimension on the order of 10^6, a time dimension on the order of 10^2, and a limited number of subjects on the order of 10^2. Nonlinear KCCA or classification problems involve estimating millions of unknowns with very few reference points. This imbalance can cause instability and overfitting in nonlinear methods. For example, in classification tasks, prior studies have shown that Support Vector Machines perform well with linear kernels, while nonlinear kernels are rarely used (Wang, 2009; Wang et al., 2007). Similar issues have been observed in decision tree-based models (Langs et al., 2011) and deep CCA (Xifra-Porxas et al., 2021). For the activation detection problem, when the activation pattern is treated as an unknown parameter with dimensionality equal to the number of voxels, even linear kernel methods without adequate regularization tend to overfit (Yang et al., 2018). Likewise, prior analyses indicate that certain nonlinear kernels can underperform compared with baseline methods in specific cases (Han et al., 2025).

To address overfitting, traditional methods often rely on additional datasets for supervision. For instance, some studies use task-fMRI data supplemented with simulated or pseudo-human data (Sorenson & Wang, 1996). Other approaches leverage resampled resting-state fMRI data (R. R. Nandy & Cordes, 2003, 2004), a technique also extended to linear KCCA, where the regularization parameter is chosen to maximize the difference between correlations in the task and resampled resting-state fMRI data (Laird et al., 2004; Yang et al., 2018). However, maximizing this difference does not yield optimal results as model complexity grows. Another disadvantage is that external datasets can be time consuming, and performance may depend on the choice of reference data. The second strategy is data augmentation. For example, a voxel-based shuffling method was employed in constrained local CCA to select the best parameter by maximizing robustness (Zhuang et al., 2017). However, this augmentation approach has only been applied to local methods.

Instead of relying on an external dataset for supervision, this study addresses the problem by focusing on the robustness of the activation patterns generated by nonlinear mappings. Similar approaches have been proposed in computer vision and medical imaging. A notable example is contrastive learning (T. Chen et al., 2020), introduced for pretraining Convolutional Neural Networks (CNNs). In this framework, augmented data are paired with original samples to form positive pairs (similar features) and negative pairs (dissimilar features). The CNN is then trained to maximize similarity within positive pairs and minimize similarity within negative pairs. Typically, augmented and original data are fed into the model simultaneously (T. Chen et al., 2020), forming a one-step approach. However, augmentation techniques such as cropping may inadvertently remove essential information. Another drawback is that blind augmentation without considering the underlying structure can introduce large randomness. As a result, contrastive learning often requires a very large batch size. To address this issue, augmentation methods based on previous outputs have been developed (Mo et al., 2021). This two-step approach is more closely related to the method we propose. Contrastive learning has also been extended to 3D medical imaging (Z. Zhou et al., 2021). Another self-supervised learning example is the vision transformer (Dosovitskiy et al., 2020), which divides images into patches and applies patch-wise augmentation. Further improvements, such as the pyramid structure in the Swin transformer, enhance segmentation performance by capturing fine details (Liu et al., 2021). These methods have been successfully adapted to medical image segmentation (J. Li et al., 2023; Tang et al., 2022) and connectivity detection (Bedel et al., 2023; B.-H. Kim et al., 2021; Yu et al., 2022).

In this study, we adopt the principle that an effective method should preserve prediction results under appropriate data augmentation. Building on this idea, we propose a novel augmentation technique that reshuffles voxel locations while preserving spatial correlations. We further validate this framework through analyses using both simulated and task-fMRI. Our results not only demonstrate that self-supervised learning works for fMRI activation detection problem but also explore the reasons behind its effectiveness (Zhang & Ma, 2022).

The second contribution of this study is a new back-reconstruction method. Because nonlinear kernels cannot be directly inverted (Wang et al., 2007; Yang et al., 2018), prior studies have estimated voxel importance to compute sensitivity maps in classification tasks using information calculated by the derivative. This approach originates from computer vision, where it is widely used for assessing pixel importance in image classification (Rasmussen et al., 2011; Selvaraju et al., 2017; Simonyan et al., 2013; Smilkov et al., 2017; B. Zhou et al., 2016). More recently, voxel-importance methods have been adapted for nonlinear KCCA in fMRI activation detection (Han et al., 2024, 2025). While effective across-subject average, this previous derivative-based method has several drawbacks: it often produces high subject-level variance, is computationally expensive for complex kernels, and relies on approximate derivatives, which limits accuracy. To overcome these limitations, we propose a more general inverse-mapping method that improves interpretability and computational

efficiency, while reducing naturally to linear KCCA in special cases.

Finally, we address the problem of selecting appropriate kernel functions for fMRI analysis. While a strict mathematical proof is difficult to establish, several empirical approaches have been proposed. One approach is based on the convergence properties of kernel functions, with the general observation that bounded kernels often outperform unbounded ones. For example, unbounded kernels have been shown to cause failures in kernel principal component analysis (S. Y. Huang et al., 2009), whereas bounded kernels demonstrate greater robustness to noise and outliers in KCCA (Alam, 2014). Another promising direction is the use of multiple kernels. Prior work has demonstrated that mixed kernels, which combine bounded and unbounded functions, outperform single-kernel approaches in KCCA (Zhu et al., 2012). Similar strategies are widely used in computer vision, including architectures such as residual and densely connected networks (He et al., 2016; G. Huang et al., 2017). In nonlinear KCCA, a mixed hyperbolic tangent kernel—integrating linear, nonlinear, and bounded mappings—achieves state-of-the-art accuracy (Han et al., 2025). In this study, we further validate the effectiveness of these strategies from different perspectives. First, from a performance point of view, the mixed hyperbolic tangent kernel consistently achieves the highest performance across different datasets. Second, through independent data augmentation tests, such as noise contamination (Alam, 2014) or shuffling robustness, this kernel also demonstrates superior performance. Third, we analyze the eigendecomposition and find that bounded kernels, in general, capture more relevant features than unbounded ones.

This study is arranged as follows. In Section 2, we present two task-fMRI datasets and two simulated datasets used to cross-validate the method. Section 3 describes the pipeline for the nonlinear KCCA method, the self-supervised learning scheme, and the evaluation metrics. Section 4.1 summarizes the subject-level activation pattern and performance across-subject average for the two simulated and two task-fMRI datasets. Following the results, we further discuss the three main contributions of this study. The first, in Section 4.2, is the validation of our proposed self-supervised learning approach in several directions. The second, in Section 4.3, is a comparison of our newly proposed back-reconstruction method with previously published methods, as well as across different evaluation metrics. The third, in Section 4.4, compares the eigendecomposition results as further validation of the kernel selection. Section 4.5 discusses the activation overlays with regions-of-interest (ROI), which is more closely related to the neurological implications. Finally, in Section 5, we discuss the strengths of our method, comparisons with previous methods, and its limitations.

2. DATA AND SIMULATIONS

2.1. HCP dataset

Structural and functional MRI data were obtained from the Human Connectome Project (HCP) database (Van Essen et al., 2012), which includes 3T MRI data. We focus on the working memory task-fMRI study. Ninety-two male subjects aged 26–30 years were randomly selected, and the sample size is sufficient to test differences between methods. Five subjects were dropped during preprocessing due to missing information, yielding a final sample of 87 subjects. The task-fMRI data were acquired using a multiband factor of 8, TR/TE = 720/33.1 ms, FA = 52 degrees, 72 slices, spatial resolution = $2\text{mm}\times 2\text{mm}\times 2\text{mm}$, and in-plane size = 104×90. The first 15 timeframes were removed to avoid an unsaturated T1 signal.

The HCP working memory task (an n-back paradigm) includes three trial types: targets, non-targets, and lures. A target is a stimulus that satisfies the task-specific memory rule—for example, matching the item presented two trials earlier in the 2-back condition or matching the predefined cue in the 0-back condition. A non-target is a stimulus that does not meet this rule and, therefore, should not elicit a match response. A lure is a special type of non-target that closely resembles a target because it matches a nearby item in the sequence (e.g., the 1-back or 3-back item in a 2-back block), and thus has a higher likelihood of inducing a false-positive response.

As this study focuses on methodological development for activation analysis, we do not focus on gender-related activation patterns. Furthermore, since activation is evaluated at the subject level, we expect that even if such differences exist, they will not change the main results.

2.2. In-house scans

Eighteen subjects were originally consented and recruited. Two subjects with amnestic Mild Cognitive Impairment (aMCI) were excluded from the study due to incidental imaging abnormalities (meningioma and arachnoid cyst), yielding a final sample of 16 subjects. This dataset consists of eight subjects diagnosed with aMCI and eight cognitively normal controls (NC). Data acquisition was conducted with institutional review board approval using a 3T GE HDx MRI scanner equipped with an 8-channel head coil (Jin et al., 2012). The subjects in both groups were matched for age, education, and right-handedness. The

detailed age and gender information are shown in Table 1. The acquisition parameters for the echo-planar imaging (EPI) sequence were: TR/TE = 2000 ms/30 ms, parallel imaging factor = 2, slices = 25 (coronal oblique, perpendicular to the long axis of the hippocampus), slice thickness/gap = 4.0 mm/1.0 mm, in-plane resolution 96×96 interpolated to 128×128, yielding a voxel size of $1.72\times1.72\times5\text{mm}^3$. High-resolution structural images were also acquired including a standard T1-weighted image ($0.43\times0.43\times1\text{mm}^3$) and a coplanar standard T2-weighted image ($0.43\times0.43\times2.5\text{mm}^3$).

Each subject completed two episodic-memory tasks based on human faces and common objects, referred to as faces and pictures, respectively. Each episodic-memory paradigm included six periods. Each period comprised four phases: encoding, distraction, recognition, and a short instruction prompt reminding participants of the upcoming task. During the encoding phase, participants viewed seven face–occupation pairs presented sequentially, each for 3 seconds (21 seconds total), and were asked to memorize them. This was followed by an 11-second distraction phase, during which participants were instructed to respond as quickly as possible by pressing the right button when the letter "Y" appeared and the left button when the letter "N" appeared on the screen. The recognition phase then presented 14 items—7 repeats from the encoding phase and 7 novel face–occupation pairs. Participants responded with the right button if the stimulus had been seen previously and the left button if it was new. For each episodic-memory task, the total scan duration was 9 minutes and 36 seconds, yielding 288 volumes. The first 10 seconds (5 timeframes) were removed to avoid an unsaturated T1 signal.

2.3. Contrast and effective design signal

The HCP working memory task includes three different trial types: targets, non-targets, and lures. Each is encoded as a binary time series, which is then convolved with a canonical HRF to generate a design matrix represented as $\mathbf{X}\in\Re^{390\times3}$. This study uses the contrast targets minus non-targets $\mathbf{C}=[1,-1,0]^T$. This contrast isolates neural activity specifically related to working memory matching and decision processes by subtracting baseline perceptual and response-related effects that are common to both trial types.

Table 1. Age and gender information for task-fMRI dataset.

	HCP dataset (NC)	In-house scans (NC)	In-house scans (MCI)
Number of subjects	87	8	8
Female	0	4 (50%)	3 (37.5%)
Age range in years	26–30	43–68	56–66

For the in-house episodic memory scans, the paradigm includes four trial types: instruction, control, encoding, and recognition. Each event type is encoded as a binary time series and convolved with a canonical HRF, yielding a design matrix $\mathbf{X}\in\Re^{283\times4}$. Two contrasts are then defined: Encoding minus Control (EC, $\mathbf{C}=[0,-1,1,0]^T$) and Recognition minus Control (RC, $\mathbf{C}=[0,-1,0,1]^T$). The EC contrast isolates neural activity specifically related to memory formation by removing low-level perceptual and motor components shared with the control condition, while the RC contrast highlights retrieval-related activity after similar removal of baseline task demands (Jin et al., 2012).

For a specific contrast $\mathbf{C}$, the effective design signal $\mathbf{X}_{\text{eff}}$ is generated according to the method described in Cordes et al. (2012)

$$\mathbf{X}_{\text{eff}}=\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{C}\left[\mathbf{C}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{C}\right]^{-1}. \tag{1}$$

Next, we map $\mathbf{X}_{\text{eff}}$ into the linear kernel space using the equation $K_{\mathbf{X}}=\mathbf{X}_{\text{eff}}\mathbf{X}_{\text{eff}}^T$. Subsequently, we focus on the effective design signal.

2.4. General framework for simulated data generation

Besides task-fMRI, simulated fMRI also provides valuable information for the activation detection problem as the ground truth is given. We generate simulated data using

$$\begin{aligned}\mathbf{Y}_{\text{simulated}}&=\mathbf{Y}_{\text{noise}}+\rho\mathbf{Y}_{\text{signal}},\\ \left(\mathbf{Y}_{\text{signal}}\right)_q&=\begin{cases}\mathbf{X}_{\text{add}}, & \textit{for } q\in\mathbb{M},\\ 0, & \textit{for } q\notin\mathbb{M}\end{cases}\end{aligned} \tag{2}$$

where T denotes the length of time, Q represents the number of voxels, $\mathbb{M}$ is a mask that contains all activated voxels, $\left(\mathbf{Y}_{\text{signal}}\right)_q\in\Re^{T\times1}$ is equal to the added signal $\mathbf{X}_{\text{add}}$ if voxel q is activated, $\mathbf{Y}_{\text{noise}}$ is the noise considered independent to $\mathbf{X}_{\text{add}}$, and ρ is the signal strength. Numerically, $\mathbf{Y}_{\text{noise}}$ and $\mathbf{X}_{\text{add}}$ are normalized in time with a mean of 0 and a variance of 1. The choice for these parameters is dataset specific as listed in the following sections. A consistent feature across all simulations is that $\mathbb{M}$ includes approximately 10% of the total number of voxels in the brain.

2.5. fMRI simulation 1

Algorithm 1. Procedure for determining the signal strength ρ with baseline-consistent performance

Generate $\mathbf{Y}_{signal}$, $\mathbf{Y}_{noise}$, and $\mathbf{X}_{eff}$. Start with $\rho = 0.1$
while True do
  Compute the receiver operating characteristic based on the general linear model with Gaussian smoothing (Section 3.3)
  Compute the Area Under the Curve (AUC) for a false positive rate (FPR) smaller than 0.1
  if $AUC_{FPR<0.1} < 0.035$ **then**
    $\rho = \rho + 0.005$
  else if $AUC_{FPR<0.1} > 0.05$ **then**
    $\rho = \rho - 0.005$
  else
    break
  end if
end while

The key components for the first simulated fMRI dataset are listed below:

1. Mask $\mathbb{M}$: The activation mask is selected for specific brain regions. We focus on *six* bilateral Automated Anatomical Labelling (AAL) regions (Tzourio-Mazoyer et al., 2002), including the anterior cingulate cortex, precentral gyrus, inferior frontal gyrus, insula, middle frontal gyrus, and middle temporal gyrus, which are treated as active. In total, approximately 10% of voxels in the brain (or around 19% of voxels in the gray matter) are activated. Let $\mathbb{M} \in \Re^{1\times Q}$ be the mask that equals 1 for voxels belonging to chosen AAL regions and 0 otherwise (Tzourio-Mazoyer et al., 2002). An example of the activation is shown in Figure 1a.
2. Noise $\mathbf{Y}_{noise}$: The resampled resting-state fMRI data are considered noise, as it does not contain task-related information. Since the time dimension for the resampled resting-state fMRI data in the HCP dataset is higher than that for the working memory task, we select the timeframes between 301 and 690 to match the time dimension.
3. Effective design signal $\mathbf{X}_{eff}$: Given three different event types: targets, non-targets, and lures contrasts, we use the contrast targets minus non-targets $\mathbf{C} = [1, -1, 0]^T$ to generate the effective design signal.
4. Added signal $\mathbf{X}_{add}$: We assume that the real signal does not exactly match the effective design signal. In other words, $\mathbf{X}_{add}$ is not exactly the same as $\mathbf{X}_{eff}$. It is generated based on the following: first perturb the original contrast by $\mathbf{C}' = [1, -1, 0]^T + 0.1\mathcal{N}(\mathbf{0}, \mathbf{I})$, then use Eq. (1) $\mathbf{X}_{add} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{C}'\left[(\mathbf{C}')^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{C}'\right]^{-1}$. Here, $\mathcal{N}(\mathbf{0}, \mathbf{I})$ refers to the normal distribution.
5. Signal strength ρ: Along with the noise, activation mask, added signal, and effective design signal, the simulated data and ground truth are generated using Eq. (2). The signal strength ρ is computed using Algorithm 1. Specifically, starting from $\rho = 0.1$, the algorithm will evaluate the performance from the General Linear Model with Gaussian smoothing (GLM+GS) method described in Section 3.3 and compute the Receiver Operating Characteristic (ROC) curve. The value of ρ is adjusted—either increased or decreased—such that the Area Under the Curve (AUC) for False Positive Ratio (FPR) smaller than 0.1 falls within the range of 0.035 to 0.05. This algorithm ensures that the signal strength is neither too high nor too low. Similar methods have been adopted in Yang et al. (2020) to reduce the type 1 error in fMRI data analysis. The exact value is shown in Table 2.
6. Spatial smoothing: For the baseline model GLM+GS, a Gaussian smoothing filter with fixed Full Width at Half Maximum (FWHM) of 4 mm is applied to $\mathbf{Y}_{simulated}$ (Yang et al., 2018). The correlation is computed between $\mathbf{X}_{eff}$ and $\mathbf{Y}_{simulated}$ after the smoothing. For the kernel method, *seven* steerable filters generated from FWHM = 4 mm are applied to the dataset (Yang et al., 2018). Detailed information about the smoothing kernels is provided in Appendix A.1.

Compared with previous simulation methods, our approach introduces several key differences. In contrast to Yang et al. (2020), where different contrasts were applied to different brain regions, we use a single contrast with relatively small perturbations. While this sacrifices spatial heterogeneity, it ensures that each cluster receives the same signal strength, making it easier to assess the effects of kernel-based activation detection methods in a controlled setting. Compared with Han et al. (2025), we retain the sign of α to be consistent in this study. Our practical experience suggests that for relatively small nonlinear effects, preserving the sign leads to more accurate results under the same signal strength. Another important difference is that since several kernels exhibit nearly identical performance, we increase the number of simulated subjects from 20 to 50 to obtain more stable results and reduce variance in performance evaluation.

2.6. fMRI simulation 2

The simulation method is discussed in Section 2.5 and still includes specific signals that are manually added. One concern is that the added signal may differ from real signals. We consider an alternative approach for adding the signal. We still treat the noise as resampled resting-state fMRI data, which is assumed to be unrelated to the task, but modify the way the signal is added. Specifically, for the task-fMRI data, a uniform Gaussian smoothing filter with FWHM = 4 mm is applied, and contrasts using targets minus non-targets $\left(\mathbf{C}=[1,-1,0]^T\right)$ are used to generate $\mathbf{X}_{\text{eff}}$. After the smoothing filter, voxels with the top 1% correlation to $\mathbf{X}_{\text{eff}}$ are assumed to be most likely activated. Task-fMRI data from these voxels are extracted, their signals are randomly shuffled, and the shuffled signals are added to regions with the top 10% of correlations (Cordes et al., 2012). This approach aims to generate a more realistic added signal rather than one derived solely from the hypothetical HRF.

1. Mask $\mathbb{M}$: For task-fMRI, find voxels with top 10% of correlations to the effective design signal after GLM+GS as the mask. An example of the activation is shown in Figure 5a.
2. Noise $\mathbf{Y}_{\text{noise}}$: Choose resampled resting-state fMRI data with a timeframe between 301 and 690.
3. Effective design signal $\mathbf{X}_{\text{eff}}$: Use contrast targets minus non-targets $\mathbf{C}=[1,-1,0]^T$ from the HCP working memory task to generate the effective design signal.
4. Added signal $\mathbf{X}_{\text{add}}$: Choose task-fMRI from voxels with top 1% of correlations to the effective design signal after GLM+GS. As a note, the smoothing is only used to compute correlation, so the added signal is not smoothed.
5. Signal strength ρ: The signal strength ρ is computed using Algorithm 1. The exact value is shown in Table 2.
6. Spatial smoothing: Following the setup of the first simulation, GLM with GS (FWHM = 4 mm) serves as the baseline. The kernel method employs seven steerable filters constructed with FWHM = 4 mm and applied to $\mathbf{Y}_{\text{simulate}}$.

For a similar simulation, we compare our approach with Cordes et al. (2012). In the previous work, the "true activation" was defined using voxels with a statistical threshold of $p=10^{-6}$. In contrast, we define activation based on the top 1% of voxels with the highest activation values. Regarding the mask selection, we use the top 10% of voxels that exhibit the strongest correlation with the design signal. This choice ensures that the number of activated voxels is roughly the same as in our first simulation. In contrast, Cordes et al. (2012) used a two-dimensional grid structure for the mask.

2.7. One-dimensional simulation

Consider a one-dimensional (1D) system where q represents a specific location, $Q=10^4$ is the total spatial length, and $T=100$ is the total number of time points.

1. Mask $\mathbb{M}$: Randomly choose 10% of locations $q\in\mathbb{M}$.
2. Noise $\mathbf{Y}_{\text{noise}}$: Choose Gaussian white noise with mean 0 and standard deviation 1.
3. Effective design signal $\mathbf{X}_{\text{eff}}$: Define $\mathbf{X}_{\text{eff}}=\sin(t+\phi)$ with $t\in[0,\,2\pi]$ and ϕ is a uniform random number with $\phi\in(-\pi/4,\,\pi/4)$.
4. Added signal $\mathbf{X}_{\text{add}}$: $\mathbf{X}_{\text{add}}=\sin(t)$.
5. Signal strength ρ: $\rho=0.3$. Under the given signal strength, the AUC with FPR <0.1 typically ranges from 0.04 to 0.06.
6. Spatial smoothing: No smoothing is performed.

3. METHODS

3.1. fMRI preprocessing

The fMRI data were minimally preprocessed using the SPM12 package (Ashburner, 2009), including realignment, slice-timing correction, coregistration, and normalization to MNI atlas (Glasser et al., 2013). A high-pass filter with a cutoff frequency of 1/120 Hz was then applied to remove temporal drift, as commonly recommended by fMRI preprocessing software such as SPM or fMRIPrep (Esteban et al., 2019; Frackowiak et al., 2004). No spatial smoothing was performed in this stage. Finally, the data

Table 2. Summary of noise, signal strength, and mask for the simulation data.

	$\mathbf{Y}_{\text{noise}}$	ρ	$\mathbb{M}$
fMRI simulation 1	Resting-fMRI data	0.0446 ± 0.0123	Six selected AAL regions
fMRI simulation 2	Resting-fMRI data	0.2624 ± 0.1279	10% of voxels with high correlation to $\mathbf{X}_{\text{eff}}$
1D simulation	Gaussian white noise	0.3	Randomly chosen 10% of all voxels

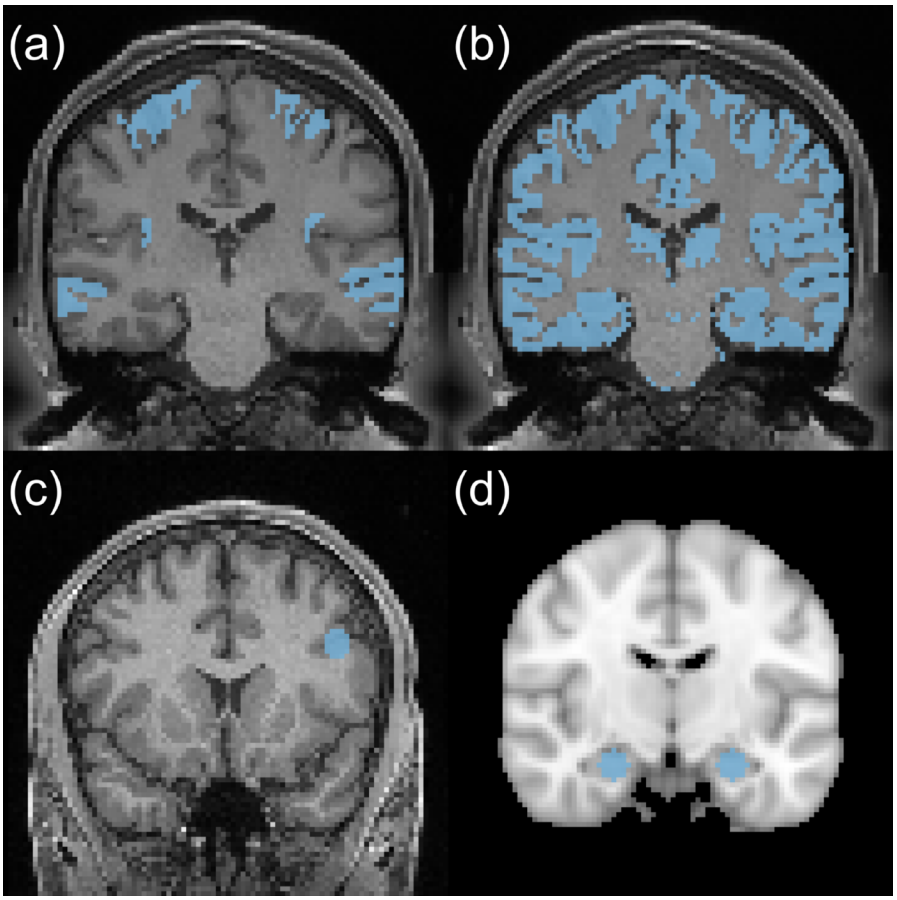


Fig. 1. Ground truth or desired activation location overlaid on the T1 image. (a) Illustration of the ground truth mask for the first simulated fMRI dataset. Activations are shown in the precentral gyrus (upper) and the middle temporal gyrus (lower). (b) Illustration of the gray matter; the subject is the same as in (a). (c) Illustration of the working memory regions of interest (ROI), with a spherical center located at MNI coordinates (–44, 14, 29) in the left prefrontal cortex and a radius of 6 mm, overlaid on the T1 image from one subject in the HCP dataset. (d) Illustration of the episodic memory ROI, with two spheres centered at MNI coordinates (24,–12,–20) and (–24,–12,–20) in the hippocampus, each with a radius of 6 mm, overlaid on the T1 image from one subject in the in-house scans.

were normalized to have a temporal mean of 0 and a variance of 1. We represent the preprocessed task-fMRI data as $\mathbf{Y} \in \Re^{T \times Q}$, where T denotes the length of time and Q represents the number of voxels. Although realignment (which internally performs motion correction) was applied, motion parameters were not used as regressor (Behzadi et al., 2007), as we aimed to avoid introducing high-frequency artifacts (J. E. Chen et al., 2017; Yakupov et al., 2017).

3.2. Steerable filters

To reduce spatial blurring, steerable filters are employed for all kernel-related analyses. Specifically, seven steerable filters generated from a Gaussian kernel with FWHM of 4 mm are used (Yang et al., 2018). The equally weighted summation of these seven filters is equivalent to a single Gaussian filter with FWHM = 4 mm. A detailed description of the steerable filters is provided in Appendix A.1.

3.3. General linear model with Gaussian smoothing filter

We use the General Linear Model (GLM) with a Gaussian Smoothing (GS) filter as a baseline model for all kernel methods. The Gaussian filter is selected to achieve the same level of smoothing as the steerable filters. The correlation is then evaluated between the smoothed fMRI data and $\mathbf{X}_{\text{eff}}$.

3.4. Linear and nonlinear kernels

For the fMRI signal, we first implement a linear mapping of the data from $\mathbf{Y}$ to the feature space $\tilde{\mathbf{Y}}$ by $\tilde{\mathbf{Y}} = \mathbf{YA} \in \Re^{T \times P}$, where $\mathbf{A} \in \Re^{Q \times P}$ is the spatial transformation matrix, T is the number of time points, Q is the number of voxels, and P is the number of voxels after spatial transformation. For example, standard GS can be done by one isotropic Gaussian filter function so that $P = Q$ (Friman et al., 2003).

More recent work uses 7 3D steerable filters such that $P = 7Q$ (Yang et al., 2018). After smoothing, $\tilde{\mathbf{Y}}$ is input to *eight* different kernels, as shown in Table 3, to map the data from the original space to the kernel spaces.

KCCA represents fMRI data through a matrix $K_{\mathbf{Y}} \in \Re^{T\times T}$. The traditionally used linear kernel can be written as $K_{\mathbf{Y}}\left(\tilde{\mathbf{Y}}_i, \tilde{\mathbf{Y}}_j\right) = \tilde{\mathbf{Y}}_i \tilde{\mathbf{Y}}_j^T$, where $\tilde{\mathbf{Y}}_i$ represents the smoothed fMRI data at time i, with the diagonal element $\left(K_{\mathbf{Y}}\right)_{ii} = \sum_q \tilde{\mathbf{Y}}_{iq}\tilde{\mathbf{Y}}_{iq}$ (Hardoon et al., 2004). In fMRI analysis, where noise levels are much higher than signal strengths, this operation generates large diagonal elements due to the noise self-correlation effect. This issue cannot be mitigated by centralized methods (Bengio et al., 2004; B. Chen et al., 2017). Instead of self-correlation, previous fMRI data have shown that noise correlation effects can occur at different times (Cordes & Nandy, 2007; X. Li, 2014). The noise correlation in fMRI (including self-correlation) can produce large matrix elements with no relationship to the signal, potentially making KCCA unstable.

We consider two approaches to mitigate noise correlation effects. The first approach involves computing the difference. If we define $K_{\mathbf{Y}}\left(\tilde{\mathbf{Y}}_i, \tilde{\mathbf{Y}}_j\right) = \| \tilde{\mathbf{Y}}_i - \tilde{\mathbf{Y}}_j \|^2$, the diagonal components are entirely removed by this operation. The second approach avoids high values through nonlinear mappings, such as a bounded linear function or a hyperbolic tangent function. These two approaches can be combined using a mixed hyperbolic tangent kernel. An additional advantage of the hyperbolic tangent function is its ability to revert to a linear kernel. We later observe that this property enhances stability across different tasks and datasets, making it a more robust choice for fMRI analysis.

We also examine the effect of increasing the noise self-correlation effect in matrix elements, such as introducing a parabolic kernel constructed by taking another square over a linear kernel. This further increases the difference between the diagonal and off-diagonal elements.

For other kernels, such as Gaussian or inverse square root (Fasshauer, 2007; Salazar et al., 2024), which map the matrix elements inversely, our findings indicate that although they roughly maintain the matrix elements within a similar range and their mapping functions can mitigate the large positive values generated by noise self-correlation, they generally do not yield significant improvements over the linear kernel.

In summary, we choose eight different kernels as defined in Table 3. We divide these kernels into *four* groups based on the function convergence and correlation properties with respect to the input. Group 1: Linear and parabolic kernels. Both are unbounded for positive and negative infinity. Group 2: Gaussian kernel and inverse square root kernel. Both are bounded in the positive region but cannot be approximated by the linear kernel. In the latter, we will abbreviate the inverse square root kernel as the inverse kernel. Group 3: Bounded linear and square kernels. These are not fully bounded functions but are designed specifically to mitigate the noise self-correlation effect. Group 4: Hyperbolic tangent and mixed hyperbolic tangent kernels. These are fully bounded functions for both positive and negative regions. We abbreviate them as tanh and mixed tanh in the latter.

All kernels in Table 3 can be generated using linear and square kernels. For these two kernels, the data are normalized to their mean absolute values to ensure consistent data ranges. A detailed formula is shown in Appendix A.2. After nonlinear mapping, the matrix value range can change, particularly the parabolic kernel or hyperbolic tangent kernel when b is small. We perform an additional normalization to ensure consistent data ranges, which improves numerical stability when computing correlations

Table 3. Summary of eight kernels, parameter constraints, and optimization steps.

Kernels	Expression	Parameters	Steps
Linear	$K_{\mathbf{Y}}\left(\tilde{\mathbf{Y}}_i, \tilde{\mathbf{Y}}_j\right) = \tilde{\mathbf{Y}}_i \tilde{\mathbf{Y}}_j^T$	$\gamma \in \left(0.1,\ 10^4\right)$	15
Parabolic	$K_{\mathbf{Y}}\left(\tilde{\mathbf{Y}}_i, \tilde{\mathbf{Y}}_j\right) = \left(\tilde{\mathbf{Y}}_i \tilde{\mathbf{Y}}_j^T + b^2\right)^2$	$b^2 \in (0,\ 1)$ and $\gamma \in \left(0.1,\ 10^4\right)$	30
Gaussian	$K_{\mathbf{Y}}\left(\tilde{\mathbf{Y}}_i, \tilde{\mathbf{Y}}_j\right) = \exp\left(-\| \tilde{\mathbf{Y}}_i - \tilde{\mathbf{Y}}_j \|^2 / \sigma^2\right)$	$\sigma^2 \in (0.1,\ 10)$ and $\gamma \in \left(0.1,\ 10^4\right)$	30
Inverse	$K_{\mathbf{Y}}\left(\tilde{\mathbf{Y}}_i, \tilde{\mathbf{Y}}_j\right) = 1/\sqrt{\| \tilde{\mathbf{Y}}_i - \tilde{\mathbf{Y}}_j \|^2 + b^2}$	$b^2 \in (0.1,\ 1)$ and $\gamma \in \left(0.1,\ 10^4\right)$	30
Bounded Linear	$K_{\mathbf{Y}}\left(\tilde{\mathbf{Y}}_i, \tilde{\mathbf{Y}}_j\right) = \min\left(C, \tilde{\mathbf{Y}}_i \tilde{\mathbf{Y}}_j^T\right)$	$C \in (0.1,\ 10)$ and $\gamma \in \left(0.1,\ 10^4\right)$	30
Square	$K_{\mathbf{Y}}\left(\tilde{\mathbf{Y}}_i, \tilde{\mathbf{Y}}_j\right) = \| \tilde{\mathbf{Y}}_i - \tilde{\mathbf{Y}}_j \|^2$	$\gamma \in \left(0.1,\ 10^4\right)$	15
Tanh	$K_{\mathbf{Y}}\left(\tilde{\mathbf{Y}}_i, \tilde{\mathbf{Y}}_j\right) = \tanh\left(b\tilde{\mathbf{Y}}_i \tilde{\mathbf{Y}}_j^T + c\right)$	$b, c \in (-1,\ 1)$, and $\gamma \in \left(0.1,\ 10^4\right)$	60
Mixed Tanh	$K_{\mathbf{Y}}\left(\tilde{\mathbf{Y}}_i, \tilde{\mathbf{Y}}_j\right) = \tanh\left(b_1 \tilde{\mathbf{Y}}_i \tilde{\mathbf{Y}}_j^T + b_2 \| \tilde{\mathbf{Y}}_i - \tilde{\mathbf{Y}}_j \|^2 + c\right)$	$\{b_1, b_2, c\} \in (-1,\ 1)$ and $\gamma \in \left(0.1,\ 10^4\right)$	200

in the kernel space. We can show that the hyperbolic tangent can approximate a linear function when b is small and $c = 0$, and the mixed hyperbolic tangent can approximate a linear or square function when b_1 or b_2 is small and $c = 0$. Theoretically, a bounded linear kernel can also approximate a linear function when C is large, but due to the highly unbalanced matrix elements, this relationship is hard to observe in our numerical study.

In nonlinear KCCA, we keep $K_{\mathbf{X}}$ as a linear kernel for two reasons. First, the nonlinear mapping can in principle be applied either to the observed fMRI data or to the hypothesized HRF regressors. In this study, we apply the nonlinear transformation only to the data side while keeping the HRF regressors linear. This choice follows a common practice in machine learning, where nonlinear feature mappings are typically applied to the input data rather than to the target variables. For example, in neural network models such as CNNs, nonlinear transformations are introduced to the data features, whereas the labels remain unchanged. Second, from a mathematical perspective, applying nonlinear mappings to both sides is often equivalent to applying a single nonlinear transformation while keeping the other side linear. Therefore, following the conventional practice, we introduce the nonlinear mapping only to the data in the present study.

3.5. Compute correlation in the kernel space

The flow map for the fMRI signal is shown in Figure 2a. Using KCCA, the eigenvectors $\mathbf{v}_{\mathbf{X}}$ and $\mathbf{v}_{\mathbf{Y}}$ are found to maximize the canonical correlation $r = \mathrm{corr}(K_{\mathbf{X}}\mathbf{v}_{\mathbf{X}}, K_{\mathbf{Y}}\mathbf{v}_{\mathbf{Y}})$ in the feature space. To avoid overfitting, we use

$$r = \max_{\mathbf{v}_{\mathbf{X}},\mathbf{v}_{\mathbf{Y}}} \frac{\mathbf{v}_{\mathbf{X}}^T K_{\mathbf{X}} K_{\mathbf{Y}} \mathbf{v}_{\mathbf{Y}}}{\sqrt{\mathbf{v}_{\mathbf{X}}^T \left(K_{\mathbf{X}}^2 + \gamma K_{\mathbf{X}}\right)\mathbf{v}_{\mathbf{X}}} \sqrt{\mathbf{v}_{\mathbf{Y}}^T \left(K_{\mathbf{Y}}^2 + \gamma K_{\mathbf{Y}}\right)\mathbf{v}_{\mathbf{Y}}}}, \tag{3}$$

where γ is a regularization parameter.

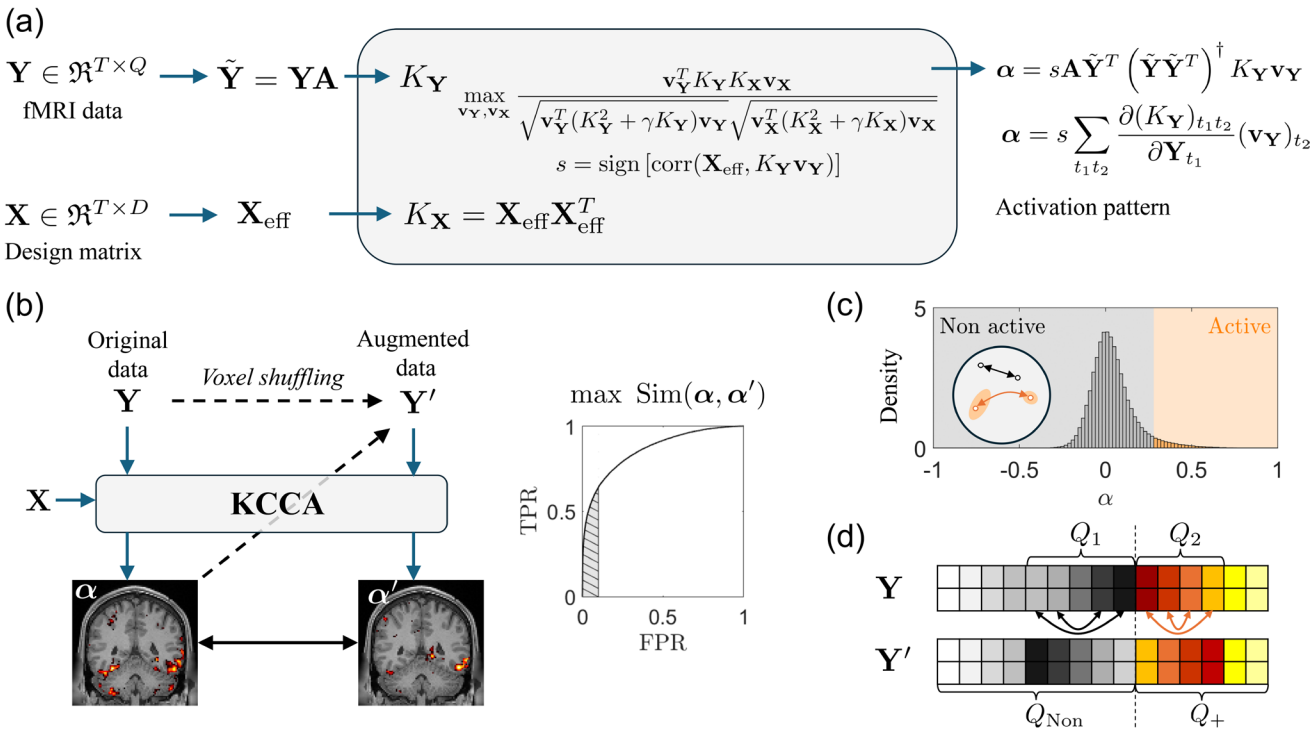


Fig. 2. (a) Flowchart for nonlinear kernel-based fMRI activation detection: The fMRI data (upper row) are mapped into kernel space using the kernels listed in Table 3. The design signal (lower row) is mapped into kernel space using a linear kernel. Correlation is computed using KCCA. Once the relationship is established in the kernel space, back-reconstruction methods, as defined in Section 3.6, are applied to extract the activation pattern. (b) Self-supervised learning framework: The original data $\mathbf{Y}$ is input into KCCA to obtain the activation map α. After voxel shuffling, the augmented data are labeled as $\mathbf{Y}'$, corresponding to an activation map α'. The ROC curve, shown on the right, is used to measure the similarity between α and α', with similarity quantified as the area under the curve (AUC) for a false positive ratio smaller than 0.1. (c, d) Illustration of voxel shuffling: (c) Distribution of α, computed using the linear kernel from the HCP dataset described in Section 4.1.3. The distributions are similar across different subjects, kernels, and back-reconstruction methods. (d) Visualization of the voxel shuffling process. The top 10% of voxels are labeled as activated. With Q_{Non} and Q_+ representing the number of non-activated and activated voxels, respectively, two clusters with voxel counts Q_1 and Q_2 are selected near the decision boundary. Shuffling is then performed within each cluster following the method outlined in Section 3.7.

3.6. Back-reconstruction methods

KCCA maximizes the correlation in the kernel space, but what we want is the activation in the original space $\alpha \in \Re^{Q\times 1}$. From dimension matching, we assume

$$\mathbf{Y}\alpha \equiv sK_{\mathbf{Y}}\mathbf{v}_{\mathbf{Y}}, \tag{4}$$

where $s = \text{sign}\left[\text{corr}\left(\mathbf{X}_{\text{eff}}, K_{\mathbf{Y}}\mathbf{v}_{\mathbf{Y}}\right)\right]$ measures the sign of the correlation between $\mathbf{X}_{\text{eff}}$ and the fMRI signal in kernel space.

The intuition behind this assumption is as follows: the correlation in the original space is written as the scalar product $\langle \mathbf{Y}\alpha \,|\, \mathbf{X}\beta \rangle$. On the left, the fMRI data are modulated by the activation pattern α and summed over spatial dimensions, while on the right, the design signal is modulated by a contrast-dependent vector β. In the kernel space, this relationship is represented by $\langle K_{\mathbf{Y}}\mathbf{v}_{\mathbf{Y}} \,|\, K_{\mathbf{X}}\mathbf{v}_{\mathbf{X}} \rangle$. By linking the representations in the original and kernel spaces, we directly arrive at Eq. (4). We propose two solutions for the above equation.

First, we insert an identity matrix $\mathbf{I} = \tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T\left(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T\right)^{\dagger}$ on the right-hand side of Eq. (4) and use the relation $\tilde{\mathbf{Y}} = \mathbf{YA}$. After canceling $\mathbf{Y}$ on both sides, we obtain

$$\text{BC 1 (This study)}: \alpha = s\mathbf{A}\tilde{\mathbf{Y}}^T\left(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T\right)^{\dagger}K_{\mathbf{Y}}\mathbf{v}_{\mathbf{Y}}. \tag{5}$$

The complete derivation of this formula is provided in Appendix A.3.

Second, following the approach of Han et al. (2025), we define $\mathbf{Y}_t$ as the fMRI data at time t and take the derivation of Eq. (4) with respect to $\mathbf{Y}_t$. This yields the following expression for α:

$$\text{BC 2 ((Han et al., 2025))}: \alpha = s\sum_{t_1 t_2}\frac{\partial\left(K_{\mathbf{Y}}\right)_{t_1 t_2}}{\partial \mathbf{Y}_{t_1}}\left(\mathbf{v}_{\mathbf{Y}}\right)_{t_2}, \tag{6}$$

where the definition of s is consistent with BC 1. The detailed derivation is shown in Appendix A.4. In the previous work (Han et al., 2025), this formulation is referred to as the heat map or sensitivity map, which measures the contribution of each voxel to the signal in kernel space. Here, we treat $\mathbf{v}_{\mathbf{Y}}$ as a constant vector so that the back-reconstruction method reduces to the results of the linear kernel in this special case.

For the special case of the linear kernel, as proven in Appendix A.5, both methods reduce to the previously published result up to a constant factor, $\alpha \propto \mathbf{A}\mathbf{A}^T\mathbf{Y}^T\mathbf{v}_{\mathbf{Y}}$ (Hardoon et al., 2004; Yang et al., 2022). For nonlinear kernels, the two different BC methods will yield different results.

For the activation analysis as shown in Section 4.1, we focus on the new proposed BC 1 method. In the meantime, we compare the two methods under the same condition in Section 4.3. Further discussion for the comparison between the two BC methods is shown in Section 5.2.

3.7. The self-supervised learning framework

The self-supervised learning framework uses augmentation methods to generate positive pairs as supervisory examples and then constructs a loss function to maximize their similarity. From the augmentation perspective, this process can be categorized as a one-step or two-step method. The one-step method applies augmentation directly to the data without considering the underlying structure, whereas the two-step method uses the model output to guide data shuffling.

3.7.1. *One-step augmentation*

The first idea examines the activation robustness under noise injection (Alam, 2014). This method generates augmented examples from the original data with added noise. Specifically, we define $\mathbf{Y}' = \mathbf{Y} + 0.1\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is Gaussian white noise. Using $\mathbf{Y}'$, we compute a new α' for both GLM and kernel methods. The distance between α and α' is defined as their mean square error (MSE), leading to the formulation of the objective function as follows:

$$\begin{aligned}&\min_{\text{parameters}} \text{MSE} = \|\alpha - \alpha'\|^2\\ &\text{subject to parameters} \in \text{Table 3.}\end{aligned} \tag{7}$$

For practical implementation, first, when computing the loss, both α and α' are evaluated using the same kernel with the same parameters, and an additional KCCA is performed to determine α'. Second, because the process of generating α' involves random variables, statistical averaging is required. Otherwise, the optimization may become trapped in false minima caused by randomness. The results for one-step augmentation are shown in Section 4.2.4. Due to the high randomness, which introduces large statistical variance, this approach incurs a high computational burden for high-dimensional data such as fMRI. Therefore, the one-step augmentation is limited to one-dimensional simulations.

3.7.2. *Two-step augmentation*

A two-step approach with information-guided augmentation can reduce the randomness introduced during shuffling (Mo et al., 2021). The idea is that isolated activated voxels are rare, so a good method will try to maintain the prediction results even after the voxel location has changed. Specifically, the framework contains data

augmentation techniques to shuffle the voxels based on their α value and a way to measure the similarity between the original result and the result after voxel shuffling. The flow map for this method is shown in Figure 2b.

Starting with a specific kernel and given parameters, we use KCCA to obtain the activation pattern α. The distribution of α is shown in Figure 2c. Assume that the top 10% of voxels with large α values are active, and the other 90% are inactive. We denote the number of activated and non-activated voxels as Q_+ and Q_{Non}, respectively. Near the decision boundary, we select two clusters: $\mathbb{Q}_1$ and $\mathbb{Q}_2$, each with Q_1 and Q_2 number of voxels from the non-activated regions and activated regions, respectively, where the blackboard bold name is for the name of the cluster. We then reverse the order within each cluster based on their α values. For instance, for all voxels in $\mathbb{Q}_1$, we switch the location of the voxel with the largest α with that of the smallest α. The same switches are performed on the voxels with the second largest/smallest α, and so on. This shuffling is repeated for cluster $\mathbb{Q}_2$. The entire process is depicted in Figure 2d. Note that shuffling is performed before smoothing.

After shuffling, we obtain a new configuration labeled $\mathbf{Y}'$ and compute the activation pattern α' using the same kernel and parameters, and we implement the KCCA again to get a new $\mathbf{v}'_{\mathbf{Y}}$. By applying different thresholds to α', we define true positive (voxels still activated after shuffling) and false positive (inactive voxels becoming activated after shuffling) to generate a Receiver Operating Characteristic (ROC) curve. An example ROC curve is shown in Figure 2b. This curve, also known as the apparent ROC curve in previous studies (Zhuang et al., 2017), measures shuffling robustness. In prior fMRI neuroscience research, the metric was designed to minimize type 1 errors to reduce the number of false positives (Skudlarski et al., 1999). Similarly, we focus on the voxels that the model strongly predicts as activated, specifically those with relatively large α'. We use the Area Under the ROC Curve (AUC) computed with a False Positive Rate (FPR) of less than 0.1 to measure the robustness. We repeat this process to find the best parameters that maximize the AUC.

Even though the two-step augmentation is based on previous outputs and the randomness is reduced, the specific design may not necessarily yield better representations of the data. In other words, activation patterns that exhibit high shuffling robustness may not overlap with gray matter well, particularly for relatively complicated kernel mappings. This problem is later referred to as reduced cross-task performance. Practically, we found that using multiple shuffling methods and averaging the results enhance the robustness of self-supervised learning. Consequently, we implement two shuffling methods. The first shuffling method permutes voxels across the entire region by setting $Q_1 = Q_{\text{Non}}$ and $Q_2 = Q_+$, while the second shuffling method permutes voxels near the decision boundary by setting $Q_1 = Q_2 = 0.5Q_+$. The AUC values obtained from the two shuffling methods are combined with equal weights. Following the convention of minimizing the objective function, a negative sign is applied to define the final objective.

$$\min_{\text{parameters}} \mathcal{L} = -\text{AUC}_{\text{FPR}<0.1}(\text{Shuffling 1}) - \text{AUC}_{\text{FPR}<0.1}(\text{Shuffling 2}), \text{subject to parameters} \in \text{Table 3}. \quad (8)$$

We use the MATLAB Surrogate Optimizer to minimize $\mathcal{L}$ (Gutmann, 2001). This optimization is generally a nonlinear problem. Because we employ a surrogate-based derivative-free solver, we adopt a fixed number of optimization steps for each kernel, as summarized in Table 3.

For the bounded linear, hyperbolic tangent, and mixed hyperbolic tangent kernels, we initialize the optimization using the parameters obtained from the linear kernel. For other kernels, we use randomly selected initial conditions as specified in Table 3. Once the optimal parameter combination is identified, we evaluate the activation using the method described in Section 3.8.

The rationale for the self-supervised learning is further validated in Section 4.2. In brief, Section 4.2.1 shows that spatial correlation is preserved after voxel shuffling, while Sections 4.2.2 and 4.2.3 further validate the self-supervised learning assumptions at the subject level. Section 4.2.4 presents results for one-dimensional simulated data with one-step augmentation. Section 4.2.5 compares performance obtained using a single shuffling method—specifically, the first shuffling method, corresponding to including only the first term in Eq. (8)—with performance obtained using both shuffling methods, in which the first and second shuffling methods are combined and both terms in Eq. (8) are incorporated. Section 4.2.6 further evaluates the convergence stability of the proposed surrogate-based optimization. By examining the relative change in the objective function under extended optimization steps, we provide empirical evidence that the selected optimization budgets are sufficient and that the optimization procedure achieves stable solutions across different kernel configurations.

3.8. Model validation

We propose five different metrics to characterize the performance of various kernels.

3.8.1. Metric 1: Ground truth

For simulated data described in Section 2.5, Section 2.6 and Section 2.7, where active voxels are known, we

generate a ROC curve by setting different thresholds for α. We use the AUC to characterize performance. In fMRI data analysis, minimizing false positives is crucial. The area for the False Positives Ratio (FPR) smaller than 0.1 is commonly used in previous fMRI publications (Skudlarski et al., 1999). We define the final accuracy as the ratio between the kernel method and the baseline model (GLM+GS):

$$r_{\text{Truth}} = \frac{\text{AUC(kernel)}_{\text{FPR}<0.1}}{\text{AUC(GLM+GS)}_{\text{FPR}<0.1}}. \quad (9)$$

This ratio is evaluated for two simulated datasets. It does not depend on the number of voxels chosen to be activated. A value $r_{\text{Truth}} > 1$ indicates that the kernel method generates better results than GLM+GS, and similarly when $r_{\text{Truth}} < 1$ shows worse performance.

3.8.2. Metric 2: Activation in gray matter vs non-gray matter

For task-fMRI data, where the ground truth is unknown, we assume that Blood-Oxygen-Level-Dependent (BOLD) activations occur primarily in gray matter (Logothetis & Wandell, 2004). The rationale includes: first, it has been shown that gray matter has three to four times the cerebral blood flow and volume compared with white matter (Gawryluk et al., 2014). Second, the BOLD signal in gray matter is three to six times larger than in white matter, with the latter also showing time delays or even negative correlations with stimulation (Schilling et al., 2023). The standard choice for HRF may not be suitable for white matter; as a result, this increases the rationale for treating white matter as false activation. Third, unlike surface-based methods, which require anatomical structures (Brodoehl et al., 2020; Mejia et al., 2020), or task denoising methods (Yang et al., 2020), our volume-based approach does not require such information, making gray matter suitable for independent testing. Fourth, this metric is consistent with earlier investigations that examined whether identified activations align with detailed gray matter profiles (Han et al., 2025; Yang et al., 2018, 2025).

Following the same idea in Metric 1, we define true positive activation in gray matter and false positive activation not in gray matter to generate the ROC curve. Gray matter is generated from the SPM package segmentation with a probability threshold set as greater than 0.5. One of the examples is shown in Figure 1b. Note that the number of voxels in gray matter is significantly larger than the number of activated voxels, which results in the AUC being close to 0.5.

Similarly, we aim to minimize false positives and reduce activation in non-gray matter, particularly in voxels that the method strongly predicts as active. We define the objective ratio for gray matter activation as follows:

$$r_{\text{Gray}} = \frac{\text{AUC(kernel)}_{\text{FPR}<0.1}}{\text{AUC(GLM+GS)}_{\text{FPR}<0.1}}. \quad (10)$$

This ratio, r_{Gray}, is evaluated for the HCP dataset and in-house scans. A large ratio indicates that the activation is more concentrated in gray matter than in other areas. This ratio does not depend on the activation threshold but does depend on the threshold for gray matter. However, we examined the output probabilities from segmentation using the SPM package, and they are concentrated around 1 or 0, leaving a few voxels on the margin.

3.8.3. Metric 3: Shuffling robustness

Similar to training accuracy in traditional machine learning, we introduce a factor to characterize shuffling robustness. As a note, we can also compute the shuffling based on GLM+GS, even if it does not have any parameters. We use $\mathcal{L}$ defined in Eq. (8) to characterize the relative accuracy between kernel and baseline model

$$r_{\text{Shuffling}} = \frac{\mathcal{L}(\text{kernel})}{\mathcal{L}(\text{GLM+GS})}. \quad (11)$$

This metric measures the robustness of certain activation patterns under voxel shuffling. It serves two purposes: characterizing our self-supervised learning method and providing an additional cross-validation of our results, especially for task-fMRI when the ground truth is not available.

3.8.4. Metric 4: Activations in regions of interest

While activation in gray matter is generally desired, as BOLD signal strength is typically higher there, it is not sufficiently specific for different tasks. This section introduces a more task-related mask to further characterize model performance.

For the working memory task, previous studies on the HCP and other working memory task-fMRI focus on activation in the left prefrontal cortex region with MNI coordinates (–44, 14, 29) (Cole et al., 2012; Lamichhane et al., 2020). The regions-of-interest (ROI) mask $\mathbb{M}_{\text{ROI}}$ is then defined as a spherical region with a 6 mm radius centered at this location, which also covers the left inferior frontal gyrus pars opercularis at (–46, 10, 26) (Rottschy et al., 2012). The mask is shown in Figure 1c.

For the episodic memory task, previous publications usually focus on hippocampal activation (Jin et al., 2012). The hippocampus plays a central role in episodic memory

encoding and retrieval, particularly through its well-established functions in associative binding, novelty detection, and pattern separation. During encoding, hippocampal circuits integrate perceptual and contextual features into conjunctive memory traces, while during recognition, they support the reinstatement and discrimination of stored representations from similar inputs. The centers of the hippocampus are defined using MNI coordinates $(24,-12,-20)$ and $(-24,-12,-20)$ (Tzourio-Mazoyer et al., 2002). For each center, we define the mask by placing a sphere with a 6 mm radius. The ROI mask $\mathbb{M}_{\text{ROI}}$ is shown in Figure 1d.

Because this region is very small, we do not use the ROC analysis proposed previously and instead compute a direct summation of the activation within this region (Lamichhane et al., 2020)

$$r_{\text{ROI}} = \sum \mathbb{M}_{\text{ROI}} \bar{\alpha}, \tag{12}$$

where $\bar{\alpha}$ is α normalized by the standard deviation, and $\mathbb{M}_{\text{ROI}}$ is the ROI mask.

3.8.5. Metric 5: Eigendecomposition

Previous analyses focused on final performance rather than the kernel itself. To characterize the information encoded in the kernel matrix, we note that all kernels listed in Table 3 are symmetric and, therefore, admit a valid eigendecomposition $K_{\mathbf{Y}} = \sum_i \lambda_i \nu_i \nu_i^T$, where λ_i is the eigenvalue and $\nu_i \in \Re^{T\times 1}$ is the corresponding eigenvector. Each eigenvector ν_i defines a direction in the implicit feature space that captures a distinct mode of data variation, with its importance measured by λ_i (Shawe-Taylor & Cristianini, 2004). In the principal component analysis problem, eigenvectors corresponding to larger eigenvalues are selected as features along the dominant directions. In real fMRI data, due to the high noise level, we find that selecting all eigenvectors, rather than examining only the ν_i associated with the largest eigenvalues, provides more stable performance across all kernels. To characterize the similarity between ν_i and $\mathbf{X}_{\text{eff}}$, we define

$$r_{\text{Corr}} = \max_{\text{parameters},i} \left|\text{corr}\left(\nu_i, \mathbf{X}_{\text{eff}}\right)\right|. \tag{13}$$

We propose this metric for several reasons. First, for a linear problem without noise, matrix $\boldsymbol{K}_{\mathbf{Y}}$ is rank 1 with $\nu_i = \mathbf{X}_{\text{eff}}$, yielding $r_{\text{Corr}} = 1$. Second, for a linear problem with noise or for a nonlinear kernel, a rigorous formulation is difficult to derive. A large r_{Corr} can indicate a high overlap with $\mathbf{X}_{\text{eff}}$. Although a rigorous connection to KCCA is implicit, a large value still indicates high flexibility of the kernel mapping. Third, we use the absolute value because the eigenvalue problem in KCCA contains squared terms for $K_{\mathbf{Y}}$ (Yang et al., 2018). As a result, sign differences do not affect the outcome.

Based on these observations, we assume that a high r_{Corr} computed with a given kernel for a specific subject indicates that the kernel encodes more relevant information, reflecting this property primarily through the kernel function itself. Practically, for a nonlinear kernel with unknown parameters listed in Table 3, we implement a grid search to find r_{Corr} across all possible mappings. As a result, it is expected that a more complicated kernel, such as the mixed hyperbolic tangent, yields a higher r_{Corr} than the hyperbolic tangent kernel. As noted, the grid search is not implemented in γ, as that is the regularization parameter for KCCA.

3.9. Statistical test

All evaluation metrics have the following properties: first, each metric yields a scalar value for each subject, with larger values indicating better performance (e.g., higher normalized activation in gray matter or ROIs). Second, the metric is normalized at the subject level, making cross-subject comparisons valid. Given two vectors of such performance scores obtained from different methods, the t-test evaluates whether the observed difference in their averages is unlikely to have occurred by chance. The null hypothesis is that the first set does not perform better than the second set. Rejecting this null hypothesis indicates statistically significant evidence that the first method achieves higher performance.

For all statistical tests listed in the study, statistical significance between the two methods was evaluated using a paired t-test across subjects. Since the analysis involved a single hypothesis test comparing the average performance of the two methods, correction for multiple comparisons was not required.

3.10. Activation threshold

Since the activation represents continuous values, we need criteria to establish the activation threshold. A commonly used approach is to resample resting-state fMRI data to compute the correlation and treat this correlation as the null distribution to define the FPR (Bullmore et al., 2001; Laird et al., 2004; R. R. Nandy & Cordes, 2003).

However, due to kernel mapping, tracking the exact values during the computation pipeline is challenging, especially when the method involves several normalization steps. In this study, we compute the null distribution by shuffling $\mathbf{v}_{\mathbf{Y}}$. Specifically, we apply random time shuffling to $\left(\mathbf{v}_{\mathbf{Y}}\right)_{\text{null}}$ and then use the randomly shuffled $\left(\mathbf{v}_{\mathbf{Y}}\right)_{\text{null}}$ as a constant to compute α_{null}. The threshold is then gradually decreased until the false activation rate reaches

0.05, analogous to the commonly used p value in previous publications. For GLM+GS, we randomly shuffle $\mathbf{X}_{\text{eff}}$ to compute α_{null}.

Compared with the previous studies, this study exhibits a relatively high FPR, because we have not changed the spatial domain, with shuffling performed only in the kernel dimension. Consequently, the computed α_{null} may yield relatively high values, necessitating a higher FPR. Nevertheless, the specific threshold value does not affect the accuracy.

4. RESULTS

4.1. Activation analysis for task-fMRI

4.1.1. fMRI simulation 1

Figure 3 shows an example of the first simulation. The activation threshold is given using $p < 0.05$. The ground-truth activated regions in Figure 3a (blue) consist of activations in both the mid-temporal gyrus (lower) and the precentral gyrus (upper). The activation pattern detected using GLM+GS is relatively small, and the linear kernel slightly improves the results but not by much. However, for all the nonlinear kernels, two clusters can be clearly seen with fine detail.

In Figure 3d and e, we present the AUC values for all the different kernels, with the gray-shaded area indicating the range for GLM+GS. For this subject, all kernels outperform GLM+GS by a relatively large margin. Specifically, for AUC with FPR < 0.1, GLM+GS achieves a value of 0.0432, which is selected between 0.035 and 0.05 to maintain consistent signal strength in the same region. The linear kernel achieves an AUC of 0.0587, while the hyperbolic tangent kernel reaches 0.0792. The other six kernels yield similar values, ranging between 0.0840 and 0.0845, indicating that these kernels provide stable and improved performance compared with GLM+GS.

Figure 4 presents the subject-wise average for this simulated dataset. In Figure 4a, the swarm plot displays the performance of all eight kernels, with exact values

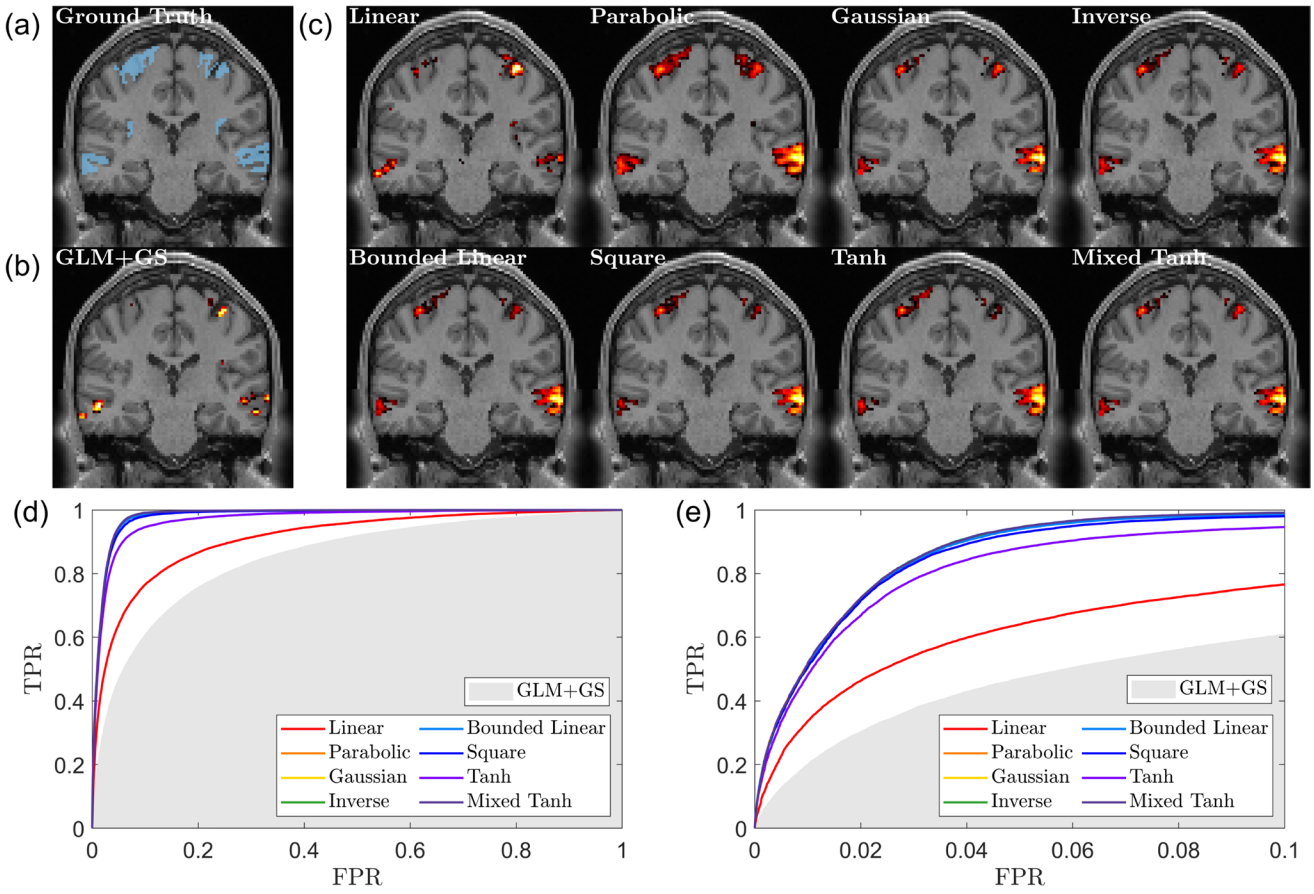


Fig. 3. Ground truth, activation map, and ROC curve for the first simulated fMRI data. (a) The ground truth, with the activation area highlighted in blue, in the T1 image at the corresponding location for reference. (b) Result from GLM+GS. The color indicates the voxel with $p < 0.05$, with a brighter color (yellow) for the larger α value. (c) Activation pattern for eight different kernels using KCCA. (d) ROC curve for different kernels. The gray shading area indicates the AUC obtained from GLM+GS. (e) A focused view of (d) for FPR < 0.1. Nonlinear kernels generate better performance than the linear kernel.

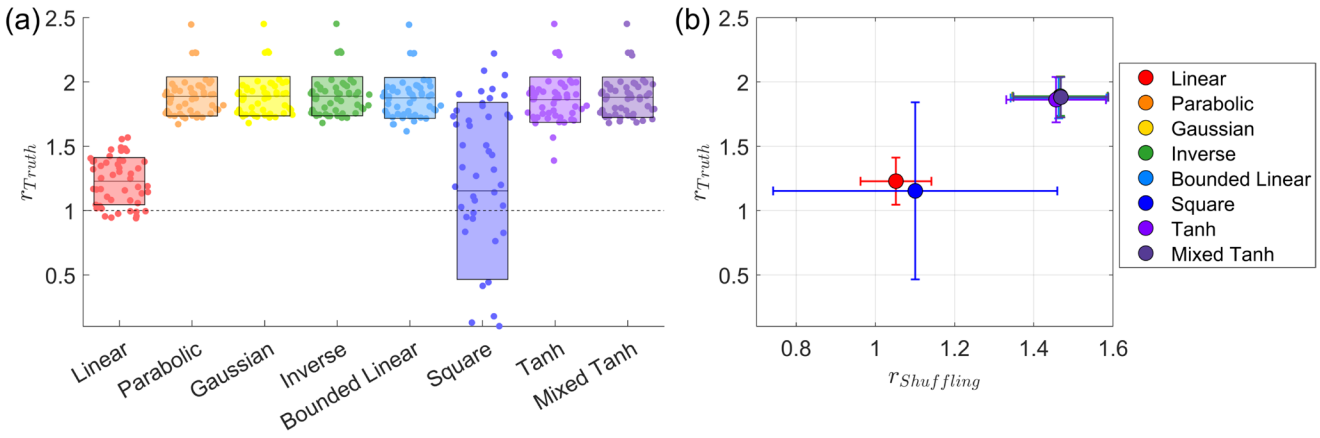


Fig. 4. Subject-wise average performance on the first simulated fMRI dataset. (a) Swarm plot of the AUC ratio evaluated using the ground truth. Results are based on 50 subjects from the first simulated dataset. Each dot represents one subject. The horizontal line and rectangle bar indicate each method's mean and variance. (b) Relationship between shuffling robustness $r_{\text{Shuffling}}$ and accuracy r_{Truth} for different kernels. Each dot represents the average over 50 subjects. In general, activations with greater robustness to shuffling also exhibit higher accuracy. Except for the square kernel, nonlinear kernels outperform linear kernels in both robustness and accuracy.

provided in Table 5. The linear kernel shows a performance increase of approximately 20%. In contrast, all other kernels—except the square kernel—achieve performance improvements of more than 80%. The square kernel performs poorly, as indicated by its large variance. Our empirical analysis suggests that, for this problem, the signal is added in a relatively uniform manner. As shown in Table 2, the signal strength is relatively low, yet the method still achieves a decent AUC value, indicating that the problem is relatively easy to learn but still requires a proper mapping function. Neither the linear nor the square kernels have parameters to tune for the mapping; their fixed mappings prevent them from adapting effectively. In contrast, the other kernels can capture the underlying structure more effectively and achieve relatively high accuracy. Another way to interpret this result is that nonlinear kernels can effectively identify optimal parameters that exhibit strong shuffling robustness. As shown in Figure 13a, with proper nonlinear kernel mapping, most subjects reach $-\mathcal{L}$ values, defined in Eq. (8), exceeding 0.18, which is close to the maximum value of 0.2 and is the highest among all datasets. Consequently, the large difference in shuffling robustness between linear and nonlinear kernels also leads to a substantial difference in accuracy.

Figure 4b presents the subject-wise average result for shuffling robustness and accuracy. We observe that the linear kernel exhibits the lowest shuffling robustness, and its average accuracy is also relatively low. In contrast, all nonlinear kernels, except for the square kernel, achieve both high accuracy and strong shuffling robustness. This result supports our assumption that the self-supervised learning approach can effectively identify the optimal parameters for nonlinear kernel mapping, ensuring both robustness and accuracy in activation detection.

4.1.2. *fMRI simulation 2*

Figure 5a presents an example from 1 of the 20 simulated subjects from the second simulated dataset. In this case, the ground-truth activation consists of relatively small clusters compared with the first simulation shown in this study. Figure 5b and c compares the performance of GLM+GS with eight different kernels. Except for the parabolic kernel, all other kernels produce activation patterns that closely resemble those of GLM+GS. The main difference is that most kernels generate relatively larger activation clusters than the ground truth. In contrast, the mixed hyperbolic tangent kernel, under the same thresholds method, produces smaller cluster sizes, effectively reducing false activations while maintaining accurate detection.

In Figure 5c and d, we present the AUC values with FPR < 0.1 for all methods. The baseline GLM+GS achieves an AUC of 0.0366, while the linear kernel improves slightly to 0.0428. The best-performing method for this subject is the mixed hyperbolic tangent kernel, which achieves an AUC of 0.0564. In contrast, the worst-performing method is the parabolic kernel, with an AUC of 0.0228. The other kernels fall within a narrow range between 0.0417 and 0.0447.

Figure 6 presents the subject-wise average for the second simulated fMRI data. In Figure 6a, the swarm plot illustrates the performance distribution across different kernels. Compared with the first simulation, the performance differences between kernels are more pro-

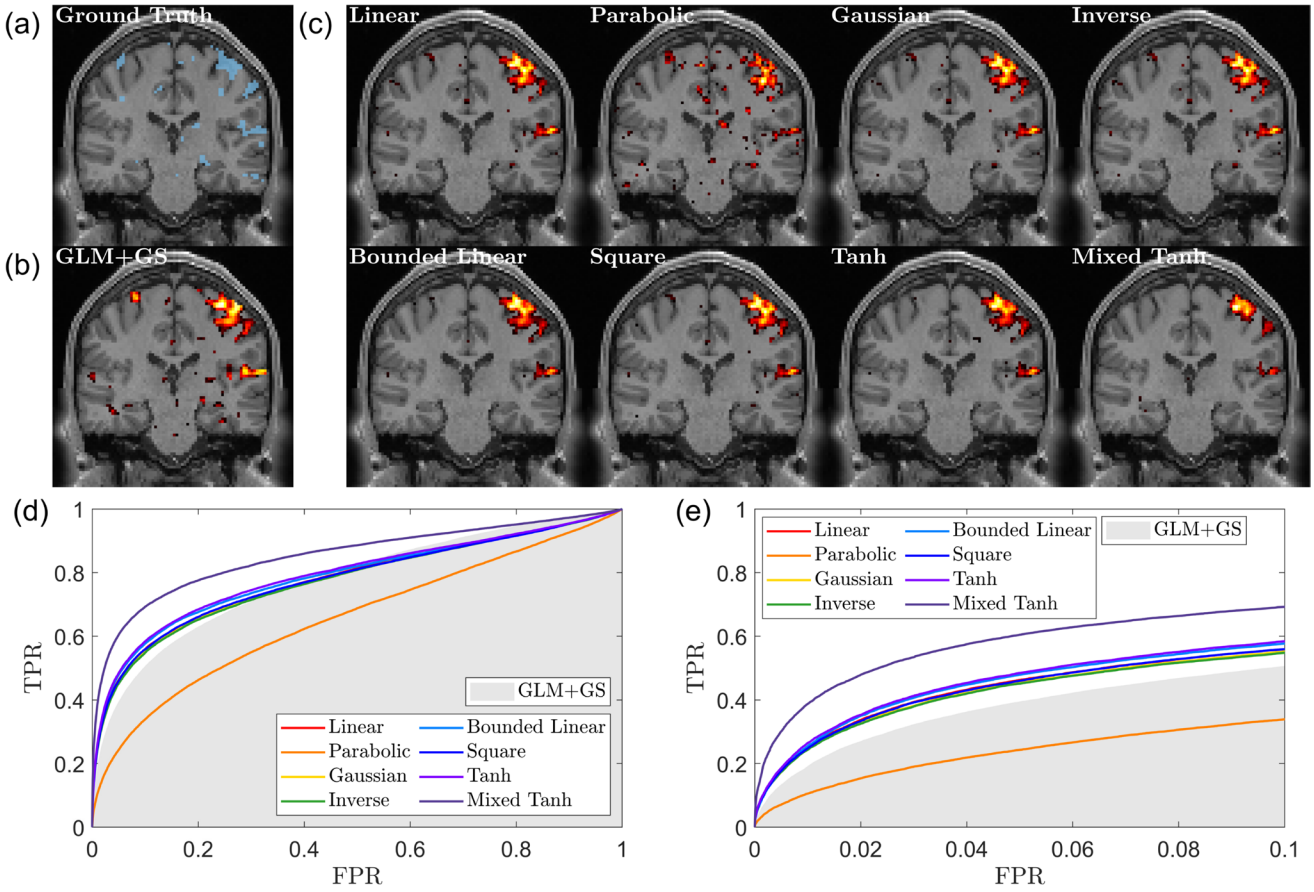


Fig. 5. Ground truth, activation map, and ROC curve for the second simulated fMRI data. (a) The ground truth, with the activation area highlighted in blue, in the T1 image at the corresponding location for reference. (b) Result from GLM+GS. The color indicates the voxel with $p < 0.05$, with a brighter color (yellow) for the larger α value. (c) Activation pattern for eight different kernels using KCCA. (d) ROC curve for different kernels. The gray shading area indicates the AUC obtained from GLM+GS. (e) A focused view of (d) for FPR < 0.1. The mixed hyperbolic tangent kernel demonstrates the highest performance, achieving the largest AUC.

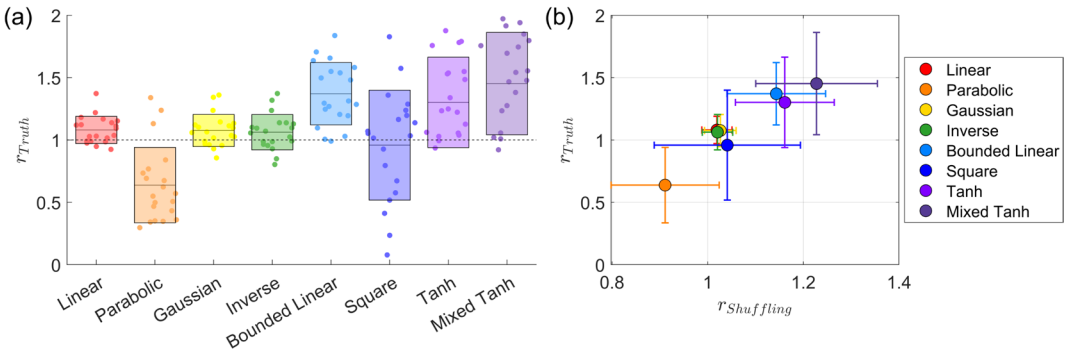


Fig. 6. Subject-wise average performance on the second simulated fMRI dataset. (a) Swarm plot for the AUC ratio evaluated using ground truth. Results based on 20 subjects taken from the second simulated fMRI dataset. Each dot represents the result from one subject. The horizontal line and the rectangle bar indicate each method's mean and variance. (b) Relationship between shuffling robustness $r_{\text{Shuffling}}$ and accuracy r_{Truth} for different kernels. Each dot represents the average over 20 subjects. In general, activation with robust shuffling also has higher accuracy. The mixed hyperbolic tangent has the maximum performance for both shuffling and accuracy.

nounced. Specifically, the linear kernel exhibits very small variance, indicating highly stable performance that closely tracks the baseline model. The Gaussian and inverse kernels show nearly identical behavior to the linear kernel. Similar to the first simulation, the square kernel performs poorly, exhibiting substantially larger variance, with many subjects falling below the baseline.

Figure 6b presents the subject-wise average ratio between shuffling robustness and accuracy. A strong linear relationship is observed between these two metrics. The mixed hyperbolic tangent kernel achieves the highest performance in both shuffling robustness and accuracy, whereas the parabolic kernel performs the worst in both cases.

4.1.3. HCP dataset

Figure 7 presents one of the selected examples, with Figure 7a showing that the gray matter serves as the reference. The results from GLM+GS are shown in Figure 7b. Results for the eight kernels resulting from the back-reconstruction methods are displayed in Figure 7c.

The activation patterns obtained from the eight kernels are approximately categorized into three groups. The first group consists of the parabolic kernel, which produces many small clusters within the brain, indicating an unreliable activation pattern. The linear, Gaussian, and inverse kernels exhibit structures that closely resemble the baseline GLM+GS model, with the primary difference being that the baseline model appears more smoothed. The bounded linear, square, hyperbolic tangent, and mixed hyperbolic tangent kernels share a similar activation structure.

Figure 7d and e displays ROC curves for the different kernels, using the gray matter as the ground truth. Since most voxels in the gray matter are inactive, the AUC values are close to 0.5. Compared with GLM+GS, indicated by the gray-shaded area, the mixed hyperbolic tangent

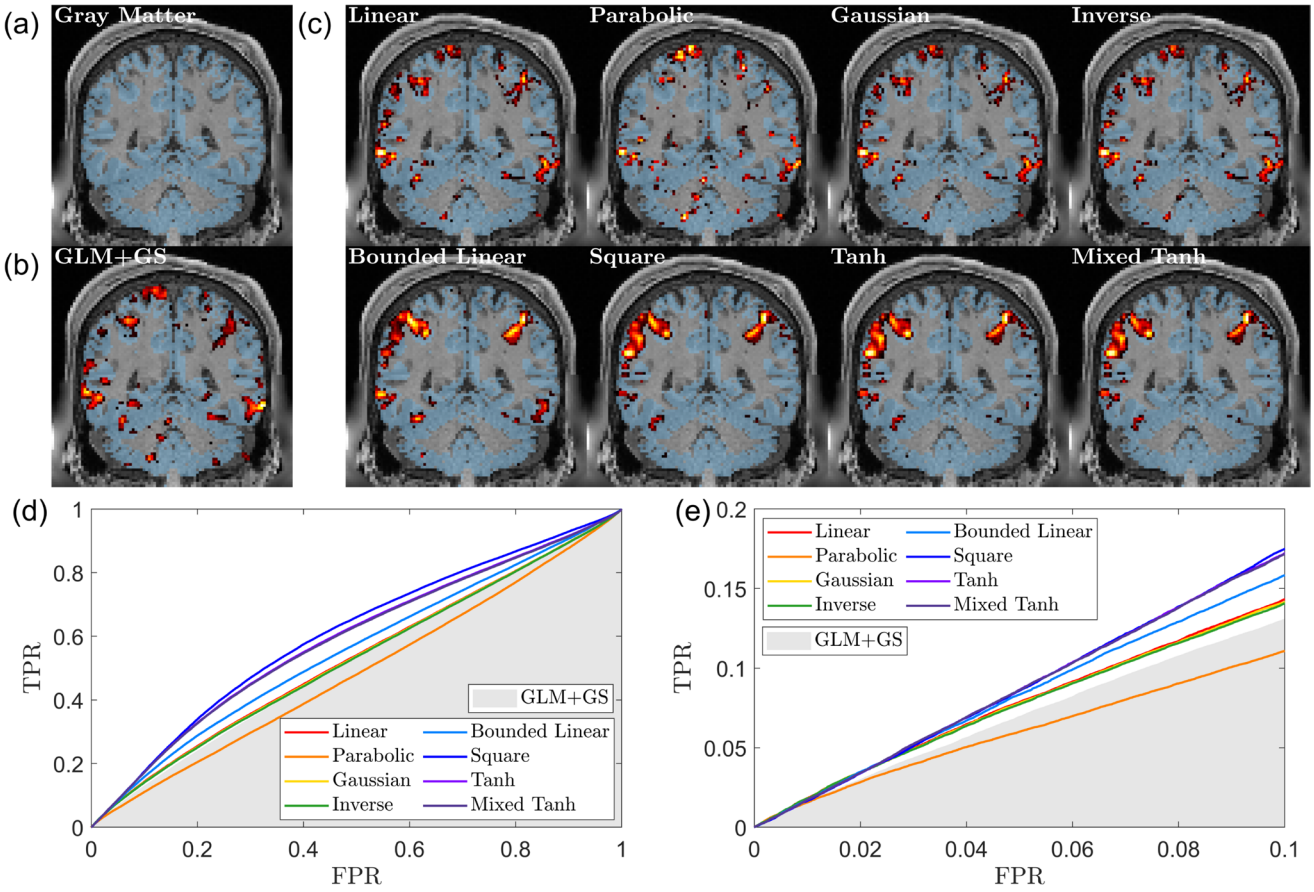


Fig. 7. Gray matter, activation map, and ROC curve for task-fMRI data (HCP dataset). (a) The gray matter (shown in blue) was determined using segmentation probabilities larger than 0.5, overlaid on the T1-weighted image. (b) Activation map obtained using GLM+GS. Voxels with $p < 0.05$ are color coded, where brighter colors (yellow) indicate larger α values; the gray matter mask (blue) and T1-weighted image are overlaid for reference. (c) Activation pattern for eight different kernels using KCCA. (d) ROC curves for the different kernels, with the gray shading area representing the AUC obtained from GLM+GS. (e) A focused view of (d) for FPR < 0.1. The parabolic kernel demonstrates poor performance, as indicated by scattered activations and the corresponding curve. In contrast, the bounded kernel such as tanh or mixed tanh demonstrates optimal performance in detecting activation in gray matter.

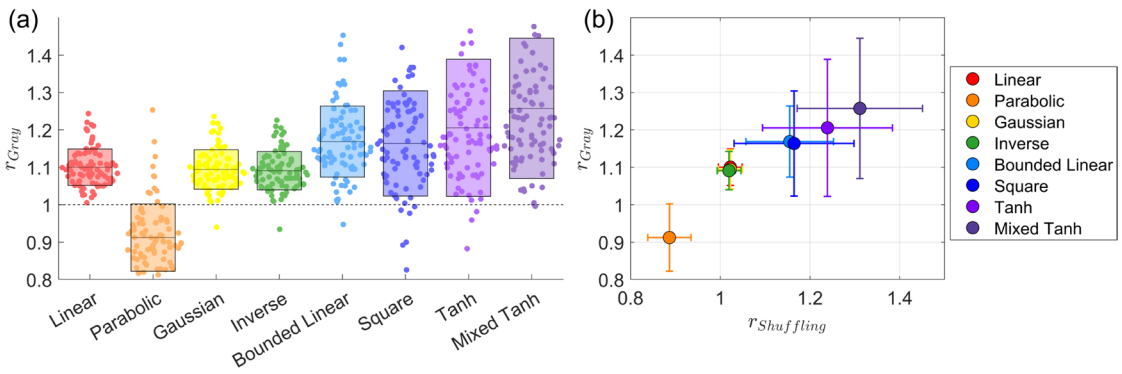


Fig. 8. Subject-wise average performance on the HCP dataset. (a) Swarm plot for the AUC ratio evaluated using ground truth. Results based on 87 subjects taken from the HCP dataset. Each dot represents the result from one subject. The horizontal line and the rectangle bar indicate each method's mean and variance. (b) Relationship between shuffling robustness $r_{\text{Shuffling}}$ and accuracy r_{Gray} for different kernels. In general, activation with robust shuffling also has higher accuracy. The mixed hyperbolic tangent has the maximum performance for both shuffling and accuracy.

kernel achieves the highest AUC and identifies a larger number of activated voxels within the gray matter. In contrast, the parabolic kernel shows poor performance, with its curves within the gray-shaded area. Comparing the exact AUC values for FPR <0.1, GLM+GS achieves 6.89×10^{-3}, while the linear kernel reaches 7.62×10^{-3}. Among nonlinear kernels, the worst performer is the parabolic kernel at 5.91×10^{-3}, whereas the best results are obtained with the square and hyperbolic tangent kernels, both at 8.63×10^{-3}. The mixed hyperbolic tangent kernel performs slightly lower at 8.61×10^{-3}. Overall, these results suggest that the square, hyperbolic tangent, and mixed hyperbolic tangent kernels outperform the others, the parabolic kernel performs the worst, and the Gaussian and inverse kernels provide no significant improvement over the linear kernel.

In Figure 8a, we present the subject-wise average results for the HCP dataset, where r_{Gray} measures the relative gray matter overlap using GLM+GS as a reference. On average, the mixed hyperbolic tangent method performs the best, showing a 25.74% increase compared with GLM+GS. In contrast, the linear kernel exhibits a 10.03% increase, indicating that our nonlinear kernels can effectively avoid activations in undesired regions compared with traditional methods. Conversely, the parabolic kernel performs the worst for the HCP dataset, with a ratio of -8.79%, indicating less gray matter overlap compared with GLM+GS. To further validate this result, in the first column of Table 4, we list the statistical tests between the mixed hyperbolic tangent kernel and the other kernels. For the HCP dataset, the mixed hyperbolic tangent kernel significantly outperforms other kernels.

In Figure 8b, we observe a strong linear relationship between $r_{\text{Shuffling}}$ and r_{Gray}, with a correlation coefficient of 0.96. This indicates that activations with high robustness to the shuffling also tend to prefer gray matter. This not only validates our assumption that self-supervised learning based on voxel shuffling is suitable for fMRI activation detection but also provides evidence using different metrics to characterize activation patterns even without knowing the ground truth.

Table 4. Statistical tests for gray matter overlap with activation for task-based fMRI.

Kernel name	HCP dataset	In-house scans
Linear	<0.001	0.009
Parabolic	<0.001	0.018
Gaussian	<0.001	0.042
Inverse	<0.001	0.003
Bounded linear	<0.001	0.001
Square	<0.001	0.024
Tanh	<0.001	0.033

One-sided paired t-test p-values comparing the gray-matter overlap metric r_{Gray} defined in Eq. (10), between the mixed hyperbolic tangent kernel and all other kernels.

4.1.4. *In-house scans*

We utilized in-house scans providing another task-fMRI dataset to evaluate the performance of nonlinear kernel methods. An example from one subject performing the faces task with the EC contrast is shown in Figure 9. For the activation pattern, Figure 9b presents the results from GLM+GS as a reference, showing some activations in the brain stem, which are likely related to cardiac pulsation and not memory activations. In contrast, Figure 9c shows that, except for the parabolic kernel, brain stem activations disappear when using kernel methods. Another notable effect is that hippocampus activations become stronger for the hyperbolic tangent and mixed hyperbolic tangent kernels.

Figure 9d and e presents the AUC values measured by activation overlays on the gray matter for this subject. As

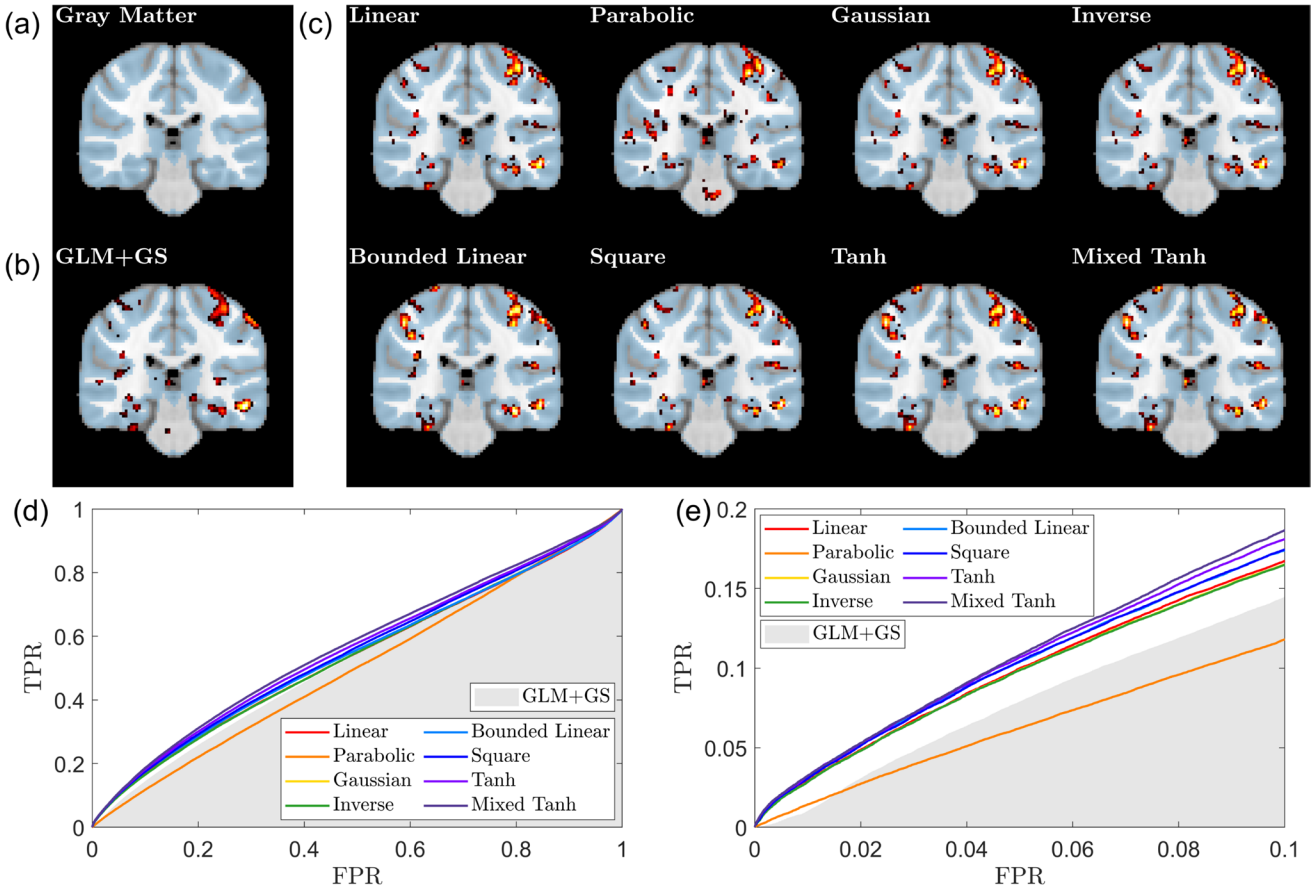


Fig. 9. Gray matter, activation map, and ROC curve for task-fMRI data (In-house scans). (a) The gray matter (shown in blue) was determined using segmentation probabilities larger than 0.5, overlaid on the T1-weighted image. (b) Activation map obtained using GLM+GS. Voxels with $p < 0.05$ are color coded, where brighter colors (yellow) indicate larger α values; the gray matter mask (blue) and T1-weighted image are overlaid for reference. (c) Activation pattern for eight different kernels using KCCA. (d) ROC curves for the different kernels, with the gray shading area representing the AUC obtained from GLM+GS. (e) A focused view of (d) for FPR < 0.1. GLM+GS and the parabolic kernel show some activations in the brain stem. In contrast, the mixed hyperbolic tangent kernel demonstrates optimal performance in detecting activation in gray matter.

a reference, GLM+GS achieves a value of 0.75×10^{-2}, while the linear kernel reaches 0.96×10^{-2}. The top two performing kernels are the mixed hyperbolic tangent and hyperbolic tangent, with values of 1.05×10^{-2} and 1.03×10^{-2}, respectively. The bounded linear and square kernels perform similarly, both achieving 1.00×10^{-2}. The Gaussian and inverse kernels also show similar performance to the linear kernel, both with values of 0.94×10^{-3}. The parabolic kernel performs the worst, with a value of 0.61×10^{-2}.

Figure 10 presents the subject-wise average for all 16 subjects across 2 tasks and 2 contrasts. In Figure 10a, we show a swarm plot of the averages across all 64 cases., along with the mean and variance for each kernel method. The specific improvement ratios for AUC with FPR < 0.1 are reported in Table 5. This dataset exhibits relatively larger variance than HCP dataset. Several factors may contribute to this variance. First, we observe that GLM+GS does not perform well, usually because these subjects involve relatively large head movements or because low signal strength leads to a higher proportion of voxels exhibiting negative correlation with the effective design signal, which causes the baseline model to have relatively low gray matter overlap with activations. Further validation of this result is shown in Figure 13d for the non-normalized values. Second, part of the variance comes from the metric definition, with the baseline used as the denominator. An alternative definition is to use subtraction between the kernel and the baseline to reduce variance. However, the ratio provides a clearer interpretation of improvement, whereas subtraction only yields an area difference that does not serve a practical meaning.

To further validate the results, in the second column of Table 5, we list the statistical tests between the mixed hyperbolic tangent kernel and the other kernels for the in-house scans. We still observe a significant performance improvement measured by gray matter overlap

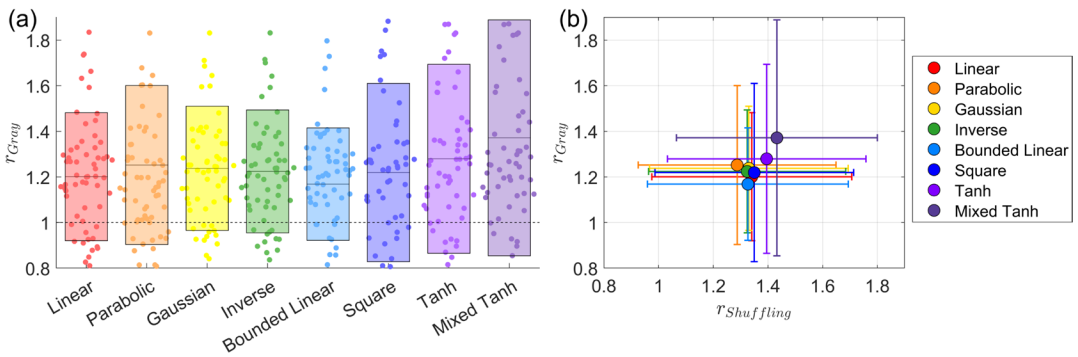


Fig. 10. Subject-wise average performance on the In-house scans. (a) Swarm plot for the AUC ratio evaluated using ground truth for the in-house scans. Each dot represents the result from one case. The horizontal line and the rectangle bar indicate each method's mean and variance. (b) Relationship between shuffling robustness $r_{\text{Shuffling}}$ and accuracy r_{Gray} for different kernels. In general, activation with robust shuffling also has higher accuracy. The mixed hyperbolic tangent has the maximum performance for both shuffling and accuracy.

with activations. However, this dataset is less robust compared with the HCP dataset.

4.2. Validation for self-supervised learning

Beyond subject-wise average accuracy, in this section, we further validate the self-supervised learning in three different directions. In Section 4.2.1, we numerically show that this method preserves the spatial correlation between voxels, unlike randomly shuffling that totally destroys it. In Section 4.2.2 we show an example from HCP dataset, and the detail process for the parameter optimization. In Section 4.2.3 we show the accuracy in subject-level analysis without normalization. In Section 4.2.4 we present a one-step argumentation technique, with application to 1D simulation. The one-step method has a major drawback that requires a large statistical average, which is difficult for high dimensional data like fMRI. Section 4.2.5 compares the performance using single shuffling (shuffling method 1) versus combined shuffling (shuffling method 1 and shuffling method 2). Section 4.2.6 further evaluates the convergence stability of the proposed surrogate-based optimization.

4.2.1. Preservation of spatial correlation

We plot the joint probability of α between certain voxels and their neighbors. Consider a voxel labeled with α_1 in Figure 11a, and its neighbor, labeled with α_2. In Figure 11b, we show the probability density for different value pairs. The distribution is computed using one selected example from the HCP dataset using the linear kernel. We randomly generated 10^6 pairs inside the brain to compute this map. The contour map is concentrated on a diagonal as the majority of voxels with similar α values tend to be neighborhoods. In other words, activated voxels tend to be a neighborhood with activated voxels, and similarly for non-activated voxels (Zhuang et al., 2017).

To characterize the distribution difference, we introduce the Kullback–Leibler (KL) divergence. Let P_{ori} denote the original distribution shown in Figure 11b, and let P_{shuffle} denote the distribution after shuffling. The KL divergence is then defined as

$$D_{\text{KL}} = \sum P_{\text{shuffle}} \log \frac{P_{\text{shuffle}}}{P_{\text{ori}}}. \tag{14}$$

Practically, a small number 10^{-6} is added in the denominator to avoid numerical overflow.

In Figure 11c, we plot the same distribution based on its new locations after shuffling. Compared with Figure 11a, the joint distribution does not change significantly, with $D_{\text{KL}} = 0.6266$, which proves that this shuffling does not destroy the spatial correlation between voxels. Finally, in Figure 11d, we plot the same distribution if we randomly shuffle the voxels in each cluster. The distribution looks totally different with $D_{\text{KL}} = 1.6159$. Moreover, our numerical results suggest that randomly shuffling the objective function is not deterministic, so the computational efficiency is not as good as with the current shuffling method.

4.2.2. Consistency during optimization

The basic assumption is that, as optimization proceeds, activations with higher robustness exhibit better accuracy. This section aims to prove this consistency during optimization.

Figure 12 shows the parameter optimization for one of the selected subjects from the HCP dataset for the mixed

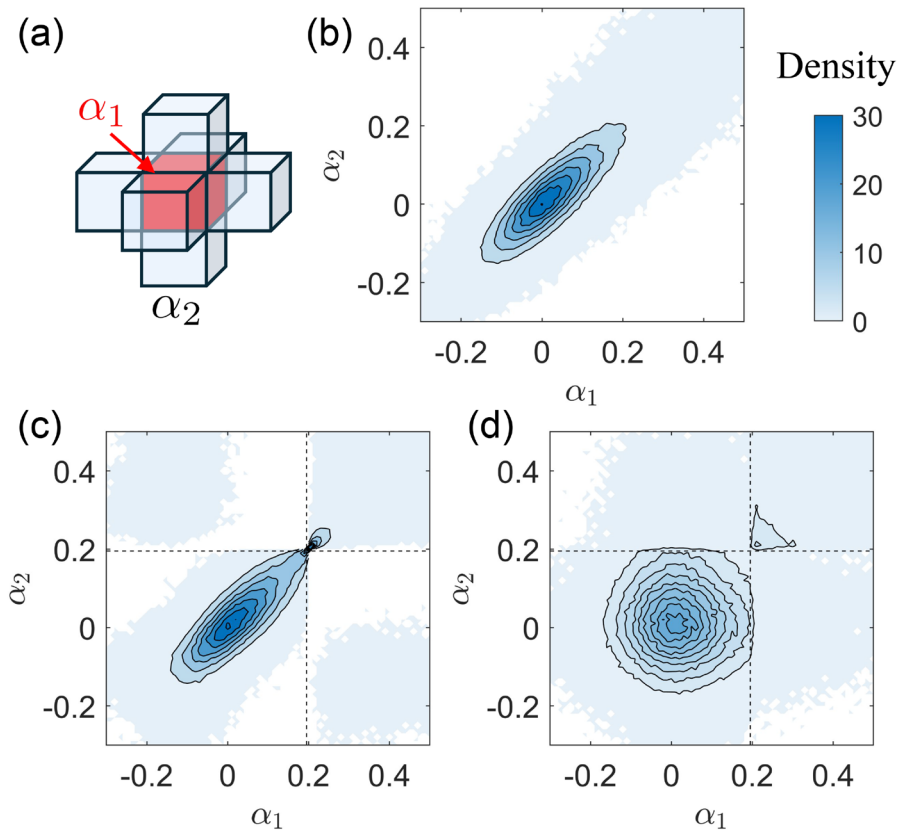


Fig. 11. Nearest neighbor distributions for α before and after voxel shuffling. The results are taken from the HCP dataset using the linear kernel. There are no significant differences between distributions across different subjects, kernels, or BC methods. (a) Illustration of the center voxel α_1 and its six neighbors α_2. (b) Original nearest neighbor distribution. A large number of voxels located on the diagonal indicate that most neighboring voxels tend to have similar α values. (c) Nearest neighbor distribution after voxel shuffling, with the dashed line indicating the decision boundary. The majority of voxels remain on the diagonal, showing that the spatial correlation structure is preserved. (d) Random shuffling with the same cluster size for comparison, where the spatial correlation structure is destroyed.

hyperbolic tangent kernel. A brief overview of our data processing pipeline is as follows:

1. Perform fMRI preprocessing without applying smoothing.
2. Select the smoothing method and kernel mapping function. We use the steerable filters in Appendix A.1, generated with FWHM = 4 mm.
3. Initialize certain parameters and run KCCA to obtain α. The initial parameters are chosen based on the results from the linear kernel: $b_1 = 10^{-3}$, $b_2 = 0$, $c = 0$, and $\gamma = \gamma_{\text{Linear}}$, where γ_{Linear} is the optimal regularization parameter found for the linear method. A small b_1 ensures that optimization starts from the linear kernel.
4. Apply voxel shuffling twice based on α to generate $\mathbf{Y}'$ and $\mathbf{Y}''$. The voxel shuffling is performed before smoothing, with the second method modifying a relatively small number of voxels.
5. For each $\mathbf{Y}'$ and $\mathbf{Y}''$, map the data to the kernel space using the same kernel and parameters. Apply the KCCA algorithm again to establish the relationship in the kernel space, then use BC 1 to obtain activation maps α' and α''.
6. Evaluate Eq. (8) by comparing the augmented activation patterns with the original activation pattern.
7. Run an optimization algorithm to find the best parameters that minimize Eq. (8).
8. Once optimal parameters are found, evaluate the activation maps using the method described in Section 3.8.2.
9. Repeat the same computations for the GLM+GS model to generate the baseline results.

In Figure 12a, we present the loss defined in Eq. (8) as a function of the optimization step. The loss curve is not guaranteed to decrease smoothly because the optimization is performed using MATLAB's surrogate optimizer rather than a gradient-based local solver. The surrogate method is a global derivative-free algorithm that alternates between constructing a surrogate model and evaluating candidate points within the bounded parameter

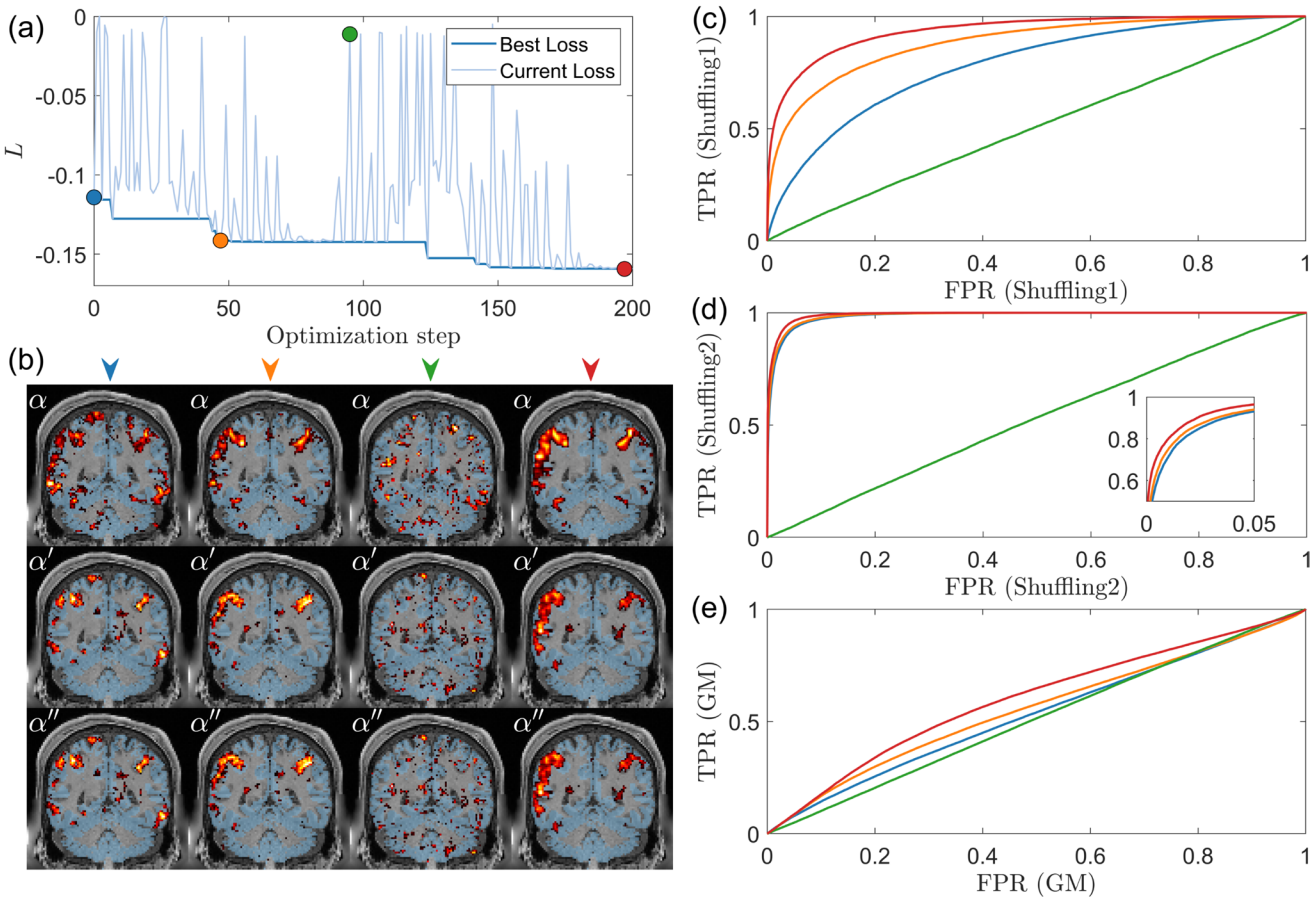


Fig. 12. Example of parameter optimization for a selected subject from the HCP dataset using the mixed hyperbolic tangent kernel. For the same subject, Figure 6 presents the performance of other kernels. (a) Relationship between optimization steps and loss defined in Eq. (8). Since the optimization is not gradient based, significant fluctuations are observed. However, the loss stabilizes after approximately 170 steps. Four different points are highlighted by blue, orange, green, and red circles, corresponding to the linear kernel, an intermediate stage with improved performance, another intermediate stage with poor performance, and the mixed hyperbolic tangent kernel achieving the minimum loss, respectively. (b) Activation pattern visualization, with bright colors indicating the top 10% of voxels with high α values, overlaid on the gray-matter mask (blue) and the T1-weighted image. α represents the original activation pattern, while α' and α'' indicate activation patterns after the first and second shuffling methods, respectively. The second shuffling method tends to produce more similar results since only voxels near the decision boundary are modified. (c) AUC curve comparison between α and α' to evaluate the effect of the first shuffling method. (d) AUC curve comparison between α and α'' to evaluate the effect of the second shuffling method. Inset: a focused view in the top-left corner. (e) AUC curve for activation in gray matter. The results clearly show that activation patterns that are more robust to shuffling also exhibit greater overlap with gray matter, confirming the assumption for self-supervised learning during the optimization process.

space. As a result, the reported objective values can fluctuate during the search, even when the algorithm is progressing toward a better solution. By monitoring the best loss, the algorithm generally converges well after 170 steps. During optimization, we select four points, indicated by blue, orange, green, and red circles, representing the initial result from the linear kernel, an intermediate stage with improved performance, an intermediate stage with poor performance, and the best result found by the optimization algorithm.

Figure 12b displays the original activation pattern α alongside the augmented activation patterns obtained from two different shuffling processes. The arrows on the top indicate the locations where the activation pattern is extracted. To ensure consistency and comparability, we plot the voxels with the top 10% of α values. Comparing the results before and after shuffling, we observe that the second shuffling method preserves more activation structure since only voxels near the decision boundary are modified. Comparing different optimization stages, we observe an increase in robustness from the linear kernel to the mixed hyperbolic tangent kernel. However, at the green-labeled location, the spreading cluster-like activation does not correspond to meaningful results, leading to low similarity between the original and augmented activation patterns.

Figure 12c and d shows the ROC curves computed using activation patterns before and after shuffling. Specifically, Figure 12c compares α and α', generated using the first shuffling method, while Figure 12d compares α and α'', generated using the second shuffling method. This type of curve has been referred to as the "apparent ROC" in previous studies (Zhuang et al., 2017). Similar to the observations in Figure 12b, the second shuffling method results in a relatively high AUC, with the lowest AUC observed at the green-labeled location and the highest AUC at the red-labeled location.

Figure 12e presents the gray matter overlap of activated voxels at each optimization stage. The results align with the shuffling robustness observations, confirming that our method successfully identifies the optimal parameter combination without requiring additional datasets.

4.2.3. Subject-level validation

For the relationship between shuffling robustness and accuracy, Figures 4, 6, 8, and 10 demonstrate that this assumption holds at the subject-wise average. Here, we examine whether it also holds at the subject level. To clarify the results, we do not apply normalization and include GLM+GS for comparison.

Figure 13 presents subject-level results for all four datasets analyzed in this study. Four kernels are selected: the linear kernel, the commonly used parabolic and Gaussian kernels, and the mixed hyperbolic tangent kernel proposed in this study. Figure 13a and b shows results for the two simulated datasets, measuring both shuffling robustness and accuracy, while Figure 13c and d displays results for the two task-fMRI datasets, evaluating shuffling robustness and activation overlays on gray matter. All methods are presented together with their mean and variance.

Figure 13a shows that in the first simulation, except for the linear kernel, all other kernels successfully capture the relationship with high accuracy and shuffling robustness. This observation aligns with Table 5, which reports normalized accuracy. In Figure 13b and c, a clearer separation between different kernels is observed, with an approximately monotonic relationship. Figure 13d presents results for the in-house scans, where GLM+GS does not perform well. Although the relationship is less pronounced, the trend remains consistent at the subject-wise average.

From a performance point of view, when comparing the two simulated datasets, it can be seen that the first simulated dataset with a nonlinear kernel performs much better than with a linear kernel. As shown in Figure 13a, the shuffling robustness $-\mathcal{L}$ defined in Eq. (8) exceeds 0.18, which is very close to the maximum value of 0.2, indicating that the nonlinear kernel can effectively capture the relationship in this case. Overall, this simulation is relatively simple, and the task remains easy to learn. Fixed mappings such as linear and square kernels lack tunable parameters and, therefore, adapt poorly, whereas nonlinear kernels better capture the underlying structure.

For the two task-fMRI datasets, minimal preprocessing was applied and explicit motion regression was not performed to avoid introducing high-frequency artifacts. To further evaluate its effect, we regressed out the six motion parameters provided in the HCP dataset and analyzed the resulting power spectra. As described in Appendix A.6, motion regression reduces power below 0.05 Hz but increases power at higher frequencies. Therefore, a trade-off exists: while motion regression can suppress motion-related components, it may also alter the spectral distribution by increasing high-frequency artifacts.

4.2.4. Comparison between one-step and two-step data augmentation

In Tables 5 and 6, we show that, in general, bounded kernels outperform unbounded kernels. Following previous work on noise contamination (Alam, 2014), we introduce an alternative augmentation approach based on noise robustness that does not require smoothing, to further provide evidence for this observation.

Using 1D simulation shown in Section 2.7, once all data are generated, the baseline model is chosen to be the GLM. Specifically, it computes the correlation at each location, represented by $\alpha \in \Re^{Q\times 1}$, and then compares it with the ground truth to generate the ROC curve. Similar to fMRI data, we conduct kernel analysis without smoothing. To simplify our results, we consider only the linear, parabolic, Gaussian, and mixed hyperbolic tangent kernels. For kernel methods without smoothing, the transformation can be interpreted as treating $\mathbf{A}$ as an identity matrix in Eq. (5).

Figure 14 presents subject-level performance for four selected kernels: linear, parabolic, Gaussian, and the mixed hyperbolic tangent. Each dot represents the result for an individual subject, with the mean and variance for each method also displayed. The $(1, 1)$ point indicates performance equivalent to that of GLM.

Figure 14 shows that for the linear kernel, nearly all results fall close to the $(1,1)$ point, indicating that in the absence of smoothing, the linear kernel behaves similarly to GLM. In contrast, the mixed hyperbolic tangent kernel has most of its points in the upper-right corner, suggesting that it improves both accuracy and robustness to noise contamination. However, unbounded kernels such as the parabolic kernel perform poorly, with most points located in the bottom-left corner.

These empirical results demonstrate that the mixed hyperbolic tangent kernel is more resilient to noise con-

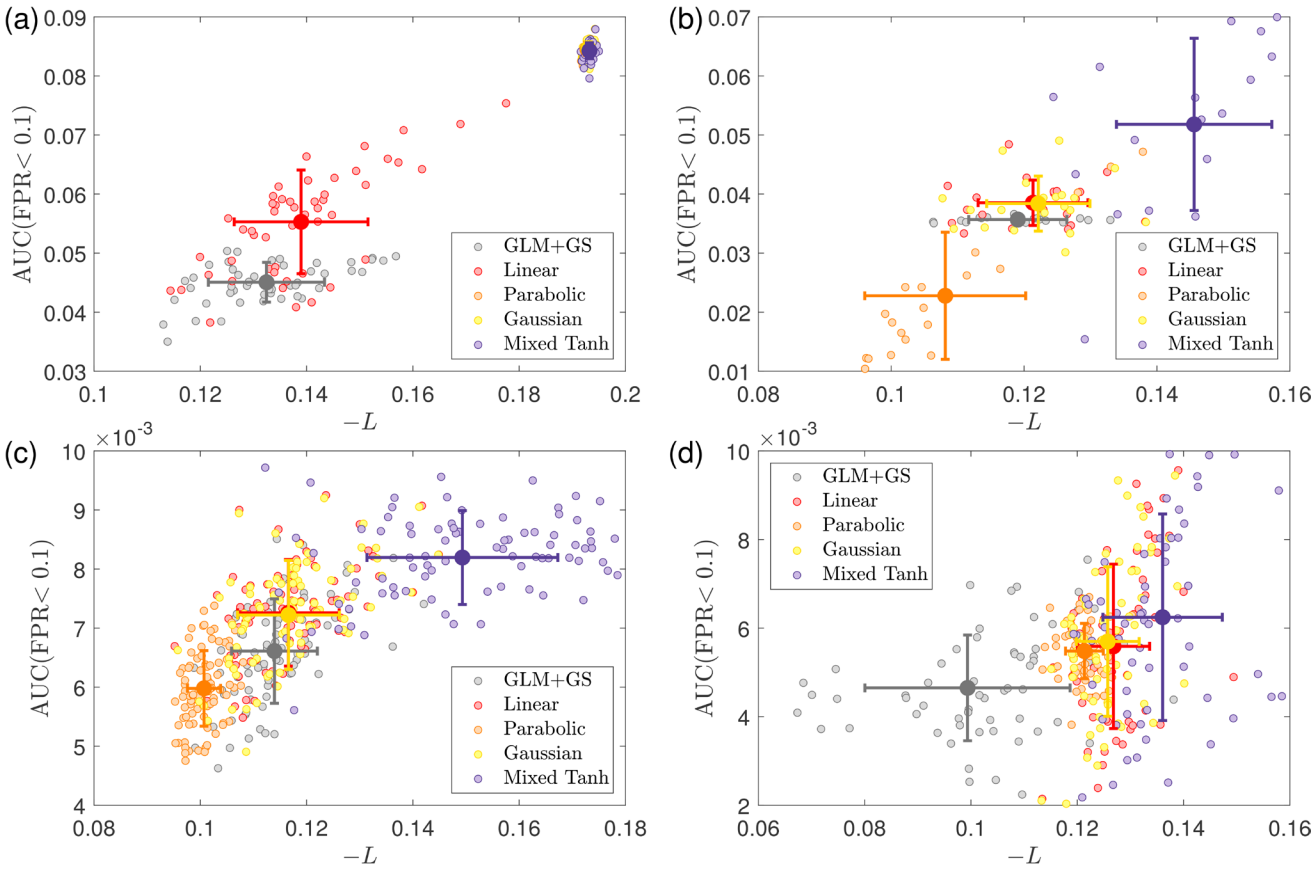


Fig. 13. Subject-level results for accuracy versus shuffling robustness, where both metrics are unnormalized. The horizontal axis represents $-\mathcal{L}$, which is the summation of two different shufflings as defined in Eq. (8). The vertical axis represents unnormalized accuracy. The values correspond to the ground truth for (a) the first simulated fMRI dataset described in Section 4.1.1 and (b) the second simulated fMRI dataset in Section 4.1.2. For (c) the HCP dataset and (d) the in-house scans, the values correspond to activation overlapping with gray matter. Error bars are included for each method.

tamination than the linear kernel, consistent with previous analyses on bounded kernels (Alam, 2014). This finding provides an alternative perspective on the kernel selection problem. However, applying this method to task-fMRI data is challenging, as it requires extensive statistical averaging during the optimization process, making it impractical for large-scale fMRI datasets.

4.2.5. *Comparison of single and combined shuffling methods*

We additionally report results obtained using only the first shuffling method. Due to computational constraints, this comparison is conducted on a subset of 40 subjects from the HCP dataset with smaller subject IDs, corresponding to approximately 50% of the full dataset, under the first back-reconstruction method.

We compare the performance of using a single shuffling method with that of using two shuffling methods. For the linear kernel, the first shuffling method achieves an accuracy of 9.86%, while two shuffling methods yield 9.84%, indicating nearly identical performance. For the mixed hyperbolic tangent kernel, using one shuffling method results in an accuracy of 24.64%, whereas using two shuffling methods improves the accuracy to 25.21%.

The primary motivation for adopting two shuffling methods is robustness. When only the first shuffling method is used, we observe two subjects whose performance drops below the GLM+GS baseline; in contrast, when both shuffling methods are employed, all subjects perform above the baseline. This suggests that incorporating multiple shuffling methods can benefit more complex kernels, such as the mixed hyperbolic tangent kernel, and improve cross-task generalization.

4.2.6. *Optimization stability analysis*

We further evaluate the stability of the surrogate-based optimization procedure. This analysis is conducted on a subset of 40 subjects from the HCP dataset with smaller subject IDs, using the first back-reconstruction method, to remain consistent with the experimental setup in Section 4.2.5.

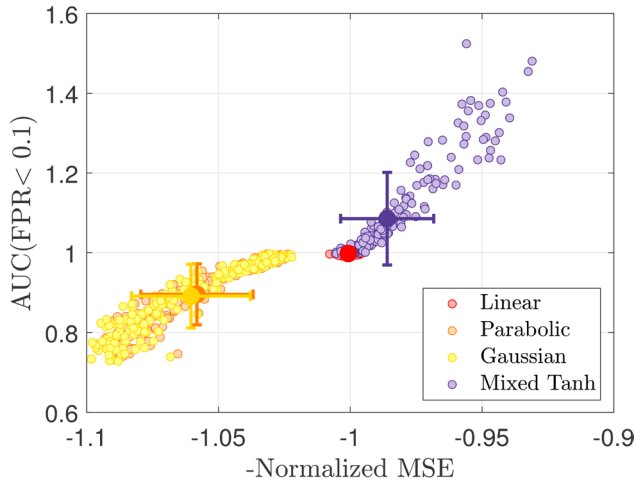


Fig. 14. Results for the 1D simulated data. The robustness to noise contamination and the accuracy of four selected kernels are evaluated without applying any smoothing. Shuffling robustness (x-axis) is quantified using Eq. (7), while accuracy is measured by the AUC under the constraint FPR < 0.1. Both metrics are normalized relative to the GLM to obtain relative ratios. Since a lower MSE indicates higher robustness, a negative sign is applied to the x-axis. Overall, increased robustness is generally associated with higher accuracy. Among the kernels considered, the mixed hyperbolic tangent kernel exhibits both the highest robustness and the highest accuracy.

To assess convergence behavior, we extend the number of optimization steps by 50% beyond the values reported in Table 3, and compute the relative change in the objective function $\mathcal{L}$. For the linear kernel (15 to 22 steps), the relative change in $\mathcal{L}$ is approximately 0.3%. For the mixed hyperbolic tangent kernel (200 to 300 steps), the relative change is approximately 0.7%.

These changes are substantially smaller than the performance differences observed across kernels (greater than 20%), indicating that the selected optimization budgets are effectively stable. This empirical evaluation supports the use of fixed optimization steps in our surrogate-based framework.

4.3. Comparison between BC methods

In this section, we compare the new BC method (denoted by BC 1) proposed in this study with the previous method, denoted as BC 2 (Han et al., 2025). For a fair comparison, both methods are applied to the same fMRI data after preprocessing. During the optimization process, each kernel in both methods utilizes the same parameter ranges specified in Table 3 and follows the same optimization steps. For the linear kernel, both BC methods produce identical results, with differences appearing only for nonlinear kernels.

Table 5 presents a comparison of accuracy with FPR < 0.1 across the four datasets analyzed in this study. For each dataset column, the best-performing kernel is highlighted in bold, while the second-best method is underlined. Overall, the BC 1 outperforms BC 2 for all the datasets, the increased ratio is more obvious for simulated dataset than for the task-fMRI dataset.

In particular, for the first simulated dataset, the new method achieves better results compared with the previous one, with performance improvements ranging from approximately 86% to 89%. Although the mixed hyperbolic tangent kernel is not numerically the best in this case, its improvement is very close to that of the best-performing kernel. For the second simulation and the two task-fMRI datasets, performance differences across kernels are more pronounced, with the mixed hyperbolic tangent kernel together with BC 1 emerging as the best method.

We also report the total AUC obtained in this study, as summarized in Table 6. In general, three observations can be made. First, the improvement rate for the overall AUC is relatively small. Second, except for the in-house scans, kernels that perform well for FPR < 0.1 also tend to show better performance for the overall AUC. Third, BC 1 still outperforms BC 2. This further demonstrates the superiority of BC 1, regardless of the activation threshold or other considerations during evaluation.

4.4. Activations within a region of interest

Activation in gray matter characterizes the overall performance, but it is not sufficiently specific for a given task. Here, we examine activation overlap with the ROI, whose definition is shown in Section 3.8.4. Because the mask is relatively small and induces a large variance, we report only the t-test results aggregated across all subjects.

In Table 7, we list four types of comparisons: kernel-based methods versus GLM+GS, as well as the mixed hyperbolic tangent kernel versus the linear kernel. We find that, except for the in-house scans comparing the linear kernel with GLM+GS, all other comparisons show statistical significance, indicating that the proposed method increases activation within the desired region.

4.5. Eigendecomposition

While the results from Section 4.1, Section 4.3, and Section 4.4 indicate the superior performance of the mixed hyperbolic tangent kernel, these metrics focus more on performance, not the kernel itself. We propose using the r_{Corr} defined in Eq. (13). The advantage is that this quantity is purely determined by the kernel and is independent of back-reconstruction or parameter optimization, which makes the comparison clearer. To narrow down our com-

Table 5. Comparison of different back-reconstruction methods.

	fMRI simulation 1		fMRI simulation 2	
Kernel name	BC 1 (This study)	BC 2 (Han et al., 2025)	BC 1 (This study)	BC 2 (Han et al., 2025)
Linear	22.86%	22.86%	8.00%	8.00%
Parabolic	88.69%	9.65%	–36.19%	**13.40**%
Gaussian	**88.91**%	12.66%	7.56%	9.94%
Inverse	88.86%	12.34%	6.22%	8.72%
Bounded linear	87.54%	11.52%	37.04%	8.31%
Square	15.32%	–6.20%	–4.16%	2.98%
Tanh	86.20%	58.64%	30.11%	–16.29%
Mixed tanh	88.10%	**66.24**%	**45.19**%	6.20%
	HCP dataset		In-house scans	
Kernel name	BC 1 (This study)	BC 2 (Han et al., 2025)	BC 1 (This study)	BC 2 (Han et al., 2025)
Linear	10.03%	10.03%	20.10%	20.10%
Parabolic	–8.79%	10.01%	25.23%	20.32%
Gaussian	9.41%	12.68%	23.79%	22.14%
Inverse	9.09%	10.12%	22.42%	22.35%
Bounded linear	16.87%	10.26%	16.85%	30.85%
Square	16.37%	16.96%	21.93%	22.84%
Tanh	20.55%	11.27%	27.95%	23.41%
Mixed tanh	**25.74**%	**24.27**%	**37.14**%	**31.14**%

The performance improvement ratios for the partial ROC with FPR < 0.1 are reported. For each dataset and each back-reconstruction method, the best-performing method is shown in **bold**, and the second-best method is shown as underlined.

Table 6. Comparison between different back-reconstruction methods.

	fMRI simulation 1		fMRI simulation 2	
Kernel name	BC 1 (This study)	BC 2 (Han et al., 2025)	BC 1 (This study)	BC 2 (Han et al., 2025)
Linear	5.25%	5.25%	1.09%	1.09%
Parabolic	15.17%	2.92%	–18.57%	**2.51**%
Gaussian	**15.20**%	3.42%	0.49%	1.65%
Inverse	15.19%	3.41%	0.15%	1.35%
Bounded linear	15.06%	3.19%	3.11%	1.18%
Square	–6.88%	–1.34%	–5.99%	–0.03%
Tanh	14.89%	9.79%	3.32%	–14.82%
Mixed tanh	15.14%	**11.90**%	**4.78**%	1.10%
	HCP dataset		In-house scans	
Kernel name	BC 1 (This study)	BC 2 (Han et al., 2025)	BC 1 (This study)	BC 2 (Han et al., 2025)
Linear	0.55%	0.55%	4.47%	4.47%
Parabolic	–4.21%	0.57%	**9.39**%	3.24%
Gaussian	0.46%	1.06%	5.51%	3.95%
Inverse	0.35%	0.66%	5.72%	4.01%
Bounded linear	2.99%	0.70%	3.60%	**6.38**%
Square	3.62%	3.86%	3.53%	1.89%
Tanh	4.54%	1.21%	4.78%	2.41%
Mixed tanh	**6.50**%	**5.34**%	5.67%	2.93%

The performance improvement ratios are defined using Eq. (9) or Eq. (10) for the total ROC. For each dataset and each back-reconstruction method, the best-performing method is shown in **bold**, and the second-best method is shown as underlined.

Table 7. Statistical test for activation within the region of interest.

Method Comparison	HCP dataset	In-house scans
Linear vs GLM+GS	< 0.001	0.1030
Tanh vs GLM+GS	< 0.001	0.0358
Mixed tanh vs GLM+GS	< 0.001	< 0.001
Mixed tanh vs linear	< 0.001	0.0089

One-sided paired t-test p-values comparing the r_{ROI}, defined in Eq. (12), among the GLM+GS, linear kernel, hyperbolic tangent kernel, and mixed hyperbolic tangent kernel.

Table 8. Statistical tests for eigendecomposition under fixed mapping complexity.

	Linear vs Parabolic	Reduced Tanh vs Linear	Reduced Tanh vs Parabolic
1D simulation	< 0.001	< 0.001	< 0.001
fMRI simulation 1	< 0.001	< 0.001	< 0.001
fMRI simulation 2	0.239	0.002	0.005
HCP dataset	< 0.001	< 0.001	0.002
In-house scans	0.532	< 0.001	< 0.001

One-sided paired t-test p-values comparing the kernel decomposition measure r_{Corr}, defined in Eq. (13), ` hyperbolic tangent kernels.

Table 9. List of the average r_{Corr} among different kernels.

	Reduced Tanh	Tanh	Mixed Tanh
1D simulation	0.8921	0.8933	0.8936
fMRI simulation 1	0.4008	0.4715	0.4886
fMRI simulation 2	0.3197	0.3836	0.4212
HCP dataset	0.2932	0.3492	0.3622
In-house scans	0.4443	0.4775	0.4947

List of the average r_{Corr}, defined in Eq. (13), among the reduced hyperbolic tangent, hyperbolic tangent, and mixed hyperbolic tangent kernel.

parison, we examine r_{Corr} in two ways: kernels with the same complexity but different convergence properties and kernels with the same convergence properties but different complexity.

We compare results among kernel mappings with the same complexity (one unknown parameter). Several details are listed as follows. First, we note that the hyperbolic tangent kernel proposed in Table 3 has two parameters. To address this, we propose a reduced tanh with only one parameter by removing the constant term: $K_Y\left(\tilde{\mathbf{Y}}_i, \tilde{\mathbf{Y}}_j\right) = \tanh\left(b\tilde{\mathbf{Y}}_i\tilde{\mathbf{Y}}_j^T\right)$. Second, the maximization in Eq. (13) is performed via grid search with 200 discretizations in each dimension using equal spacing. Third, for comparison, we also include the linear kernel. In conclusion, from a complexity perspective, both the parabolic kernel and the reduced tanh have one parameter, whereas the linear kernel has no tunable parameter. The reduced tanh is bounded, the linear kernel is unbounded but first order, and the parabolic kernel exhibits quadratic divergence.

Table 8 shows the statistical significance of the reduced hyperbolic tangent kernel over unbounded kernels such as the linear and parabolic kernels. We also observe that, in three of the five tasks, the linear kernel outperforms the parabolic kernel. Moreover, across all five tasks, the reduced hyperbolic tangent kernel shows better performance than both the linear and parabolic kernels. This indicates that the bounded kernel encodes richer information compared with unbounded kernels.

Next, we examine the effect of unknown parameters under the same boundedness property. This type of analysis is motivated by previous studies of mixed kernels (Zhu et al., 2012). We compute r_{Corr} among reduced hyperbolic tangent, hyperbolic tangent, and mixed hyperbolic tangent. Since increasing the parameters under this setup always increases the maximum value in Eq. (13), the corresponding t-test does not provide additional insight. Therefore, we instead compare the raw values to assess this effect.

Table 9 reports the r_{Corr} values for different kernels. In the 1D simulation, r_{Corr} is substantially higher, likely reflecting the relatively simple structure of this problem compared with fMRI simulations or task-fMRI data; in more complex settings, kernels with additional terms can encode richer information and thus yield higher r_{Corr}. However, whether increasing kernel complexity—and consequently r_{Corr}—consistently translates into improved activation accuracy remains unclear. This issue is further discussed in Section 5.1.4 and Section 5.3.1.

5. DISCUSSION

5.1. Self-supervised learning

5.1.1. Comparison with computer vision

Compared with the contrastive learning framework proposed in T. Chen et al. (2020), a commonality is that both methods use data augmentation techniques to generate supervisory examples. However, there are also several key differences.

First, in CNN-based classification problems, class labels are treated as the ground truth. Self-supervised learning primarily focuses on pre-training, followed by a fine-tuning stage to determine the model parameters. In contrast, our work utilizes self-supervised learning to determine unknown parameters, as the KCCA algorithm maximizes correlation in the kernel space, which is equivalent to the supervised learning approach used in previous work.

Second, in computer vision, a one-step approach—data augmentation without using model outputs—is

usually adopted (Mo et al., 2021). In our study, we primarily focus on a two-step voxel shuffling. The trade-off is that one-step augmentation is much more efficient for GPU parallelization, whereas the two-step approach avoids randomness and requires fewer statistical averages. Due to practical computational considerations, we only use the two-step method for fMRI.

Third, in computer vision, both positive and negative pairs are used (T. Chen et al., 2020). In our study, however, only positive pairs are considered, where we maximize the similarity between the original data and the augmented data. While it is possible to generate negative pairs in fMRI data, such as by shuffling the time domain $\mathbf{v}_Y$ to determine the activation threshold, we found this process requires extensive statistical averaging. As a result, we did not implement negative pairs.

Another related topic is adversarial training (Goodfellow et al., 2014), where adversarial examples are intentionally introduced during training to enhance robustness. Some adversarial examples are generated using similar augmentation techniques compared with contrastive learning (Cohen et al., 2019; Liu et al., 2019; Prakash et al., 2018). However, the current adversarial training still requires labeling examples. In other words, the loss function is evaluated using ground truth and the output from adversarial examples. In our study, we observe that maximizing the similarity between the original and augmented output is already enough for the model to generate accurate activation. Therefore, we call the method self-supervised instead of adversarial training.

5.1.2. Comparison with previous optimization techniques in an fMRI study

Compared with the voxel shuffling method used previously (Zhuang et al., 2017), our method can shuffle the entire fMRI dataset at once, making it suitable for global methods such as KCCA.

Compared with previous work that used resampled resting-state fMRI data as a reference (R. R. Nandy & Cordes, 2003, 2004; Yang et al., 2018), we found that maximizing the difference between task-fMRI and resampled resting-state fMRI data does not yield stable results for nonlinear cases. One possible reason is that nonlinear kernels introduce too many opportunities to create differences between task and resampled resting-state fMRI data, making purely numerical comparisons unstable. Our self-supervised learning approach also reduces the data used during the optimization extraction process.

An alternative approach for task-fMRI is to define the objective function based on activation overlapping with gray matter, a method similar to that used in deep CCA studies (Yang et al., 2025). However, we find that for global methods such as KCCA, maximizing activation within gray matter may yield better numerical values but does not consistently align with the expected results observed with the linear kernel. Instead, optimizing based on voxel shuffling provides a more reliable solution. Moreover, optimization via voxel shuffling is more convenient, as it does not depend on gray matter labels. The success of simulated data also indicates its broad applicability to the nonlinear KCCA problem.

5.1.3. Assumption validation

The first assumption is that we aim to create positive pairs. Therefore, we want to maintain the spatial relationship after shuffling. As shown in Figure 11, the nature of fMRI analysis is that voxels with similar activation tend to be spatially close to each other. As a result, when we shuffle the voxel locations, the order is reversed, the assumption remains valid, and the spatial correlation is maintained after shuffling.

The second assumption in self-supervised learning is that activation patterns with high robustness to shuffling also exhibit high accuracy. In other words, the assumption is about the cross-task generalization problem. From Figures 4, 6, 8, and 10, subject-wise average shows a monotonic relationship between the shuffling robustness and gray matter overlap. A similar relationship also appears in Figures 12 and 13.

We also note that the performance of self-supervised learning is closely related to the back-reconstruction method. For example, when we compute the same relationship using BC 2, the association is weaker than that observed with BC 1.

5.1.4. Strengths, limitations, and further work

The advantages of the proposed self-supervised learning method are as follows. First, from a computational perspective, it does not require an additional dataset, such as resting-state fMRI data or simulated data. Second, the method prevents overfitting across multiple simulated and task-fMRI datasets and demonstrates robustness at the subject level. Third, we show that the method is valid across different directions, offering insights into its internal mechanisms based on voxel shuffling, beyond simple accuracy improvements, which is an important aspect of self-supervised learning in computer vision (Tamkin et al., 2023).

We also note that several aspects remain unresolved. In particular, what constitutes the optimal shuffling method is still an open question. Identifying such a method is challenging due to the substantial computational cost involved. Based on our empirical observations,

we summarize several heuristic findings below. First, combining statistical averages across different shuffling methods can improve performance. Second, from a general complexity analysis, more complicated kernels benefit more from statistical averaging during shuffling. Third, we only tested combinations of different shuffling methods with equal weights, following a common first-step strategy in many self-supervised learning frameworks (e.g., Swin Transformer (Tang et al., 2022)). Although shuffling along boundaries results in smaller voxel displacements than shuffling all voxels and typically yields a slightly higher AUC, the variation in AUC across kernel mappings is small, as the optimization primarily depends on relative changes rather than absolute values. Two shuffling methods are combined with equal weights.

Second, it remains unclear whether one-step shuffling exists for the activation detection problem. For example, in Section 4.2.4, we find that augmentation without using model output exists, but it is generally hard to implement for task-fMRI due to computational limitations. One possible direction is to apply augmentation at the level of image patches (Tang et al., 2022).

Third, it is worth considering whether the current self-supervised learning approach can be extended to other problems. For example, in a typical classification-based problem, one could compute a heat map defined by voxel importance to generate supervisory examples, encouraging the model to produce consistent predictions (Rasmussen et al., 2011; Simonyan et al., 2013). This approach could also be combined with resampling techniques (Laird et al., 2004), which are typically used to break spatial correlations and generate negative pairs.

5.2. Back-reconstruction methods

5.2.1. Comparison with deep learning-based methods

Compared with deep learning methods, our approach relies on a more mathematically grounded back-reconstruction framework. Deep learning-based models, such as the Swin Transformer, also employ self-supervised learning for training (P. Kim et al., 2023; Kwon et al., 2025; Tang et al., 2022) and have the capacity to generalize across multiple tasks as foundation models (Karimi, 2024). Here, we summarize several advantages and disadvantages in comparison with deep learning.

First, in terms of design complexity, the primary source of complexity in our nonlinear KCCA framework lies in the mathematical design of the kernel map and the back-reconstruction method. Each kernel and back-reconstruction scheme has a well-defined but distinct mathematical form, and the overall performance depends strongly on these modeling choices. By contrast, in deep learning, the functional form of the model is typically fixed, while variability mainly arises from the network architecture and training configuration (e.g., network depth, width, normalization layers, learning schedules, and regularization strategies), which inevitably increases the tuning burden and may complicate interpretation.

Second, in terms of computational speed, the training phase for deep-learning-based methods typically requires expensive computational resources, such as GPUs, especially for high-dimensional fMRI data. Deep learning is also usually based on very large datasets. An advantage is that deep learning often provides faster inference once the model is trained, as no parameters need to be re-estimated. For nonlinear KCCA, a disadvantage is that parameters need to be evaluated for each subject. However, its advantages are that it requires fewer computational resources, as computation can be performed at the subject level and only CPUs are needed.

Third, while deep learning models are often equipped with post hoc interpretability tools such as saliency maps or heat maps (Selvaraju et al., 2017; Smilkov et al., 2017; B. Zhou et al., 2016), which are conceptually similar to the second back-reconstruction strategy in nonlinear KCCA, their interpretability remains limited because the relationship between the learned representation and the design signal is implicitly encoded within a complicated network architecture. In contrast, KCCA explicitly models the relationship between two domains through well-defined mathematical structures, enabling direct interpretation of how information is shared across modalities in both linear and nonlinear settings.

Another way to view interpretability is that, while deep learning demonstrates strong performance for many problems, it often suffers from issues such as covariance shift across domains, overfitting when the number of subjects is limited, or dimensionality shift. In contrast, our nonlinear KCCA maintains a clearly defined mathematical structure, making it more flexible for different problems without requiring substantial changes to the model.

5.2.2. Comparison with a previous back-reconstruction method

We give further analysis of the new proposed back-reconstruction method and the previous method (Han et al., 2025). First, for the mathematical properties, both methods can reduce to linear kernel results, as proven by Appendix A.5. Both formulas can be derived from Eq. (4), as shown in Appendices A.3 and A.4. BC 2 has more assumptions, which are less rigorous compared with the method proposed in this study.

Second, both methods preserve the sign of the activation. This means that α obtained from the BC method can be treated as the effective correlation by the kernel method, making it generally suitable for a wide range of tasks with activated or de-activated regions. This equivalence is further supported by the similarity of activation patterns.

Third, for the computational speed, both BC methods can be converted to matrix operations with given $\mathbf{v}_Y$. The time complexity for both methods is $O(QT)$, but BC 1 is more efficient due to its simpler formula, especially for complicated kernels. Practically, other operations such as reshaping could minimize their differences. We observe approximately twice the efficiency for BC 1 in computation time in the back-reconstruction.

From a performance point of view, this method offers several advantages. First, subject-wise average results indicate that this method improves accuracy compared with voxel importance, as demonstrated across multiple simulated and task-fMRI datasets. Second, it exhibits lower variance than previous methods. Using the mixed hyperbolic tangent kernel on task-fMRI data as an example, in the HCP dataset, we observed only one subject with performance lower than the baseline, whereas this number was five with the previous method. For the in-house scan, the current method resulted in 12 cases below the baseline, compared with 20 with the previous method.

There are also general limitations of the new back-reconstruction method together with the global method. First, similar to Han et al. (2025), local information is missing, making the p value typically used for fMRI activation analysis difficult to define. Second, global methods such as KCCA extract relationships across the whole brain, and it is unclear how spatial diversity may affect the results, such as regional differences in hemodynamic response function (HRF) characteristics (Erhardt et al., 2012; Glover, 1999; Lindquist et al., 2009). Developing an optimal approach to address this spatial heterogeneity problem is still an open question.

5.3. Kernel selection strategies

One of the key questions is which kernel is most efficient for extracting information from fMRI data. A rigorous answer is difficult because this process involves multiple steps, including kernel selection, parameter estimation, and back reconstruction. We attempt to address this question from three perspectives: performance on simulated and task-fMRI data, augmentation robustness, eigendecomposition viewpoint on which kernels capture information related to the target signal, and insights from prior work.

5.3.1. Bounded vs unbounded kernel

The first property we observe is the distinction between bounded and unbounded kernels. Among the eight kernels analyzed in this study, the parabolic kernel performs poorly, showing the lowest accuracy in half of the tasks. The linear, Gaussian, and inverse kernels perform similarly. The bounded linear and square kernels exhibit inconsistent results. Finally, the best-performing group consists of the hyperbolic tangent and mixed hyperbolic tangent kernels. Moreover, robustness to data augmentation further supports this finding, as the activation patterns from the mixed hyperbolic tangent kernel exhibit the highest robustness across multiple datasets, regardless of whether one-step or two-step augmentation is used.

In the eigendecomposition, we found that the parabolic kernel is usually hard to encode information from the data. Specifically, under the same model complexity with one parameter, across all datasets, the parabolic kernel is composed of eigenvectors that have small correlations with the target signal. In contrast, the hyperbolic tangent kernel contains more information, exceeding the linear and parabolic kernels with statistical significance.

One possible explanation is that bounded kernels reduce the effect of noise self-correlation. In KCCA applied to fMRI, kernel mapping embeds the spatial dimension, where large diagonal elements often arise due to noise self-correlation. A bounded kernel mitigates these large elements through continuous mapping, leading to a more balanced matrix structure that facilitates information extraction. Another perspective supporting this explanation is that bounded kernels are generally more robust to noise contamination. As demonstrated in Section 4.2.4, using an alternative data augmentation technique in 1D without smoothing, the results indicate that bounded kernels, such as the mixed hyperbolic tangent, yield consistent performance in the presence of noise contamination. This noise-robustness property is consistent with previous studies (Alam, 2014; S.-Y. Huang et al., 2009).

5.3.2. Single vs mixed kernel

Another observation relates to the unique properties of the mixed hyperbolic tangent kernel. This kernel combines both linear and nonlinear terms and, in certain special cases, can reduce to a purely linear form. Under similar bounded mappings, having more terms generally leads to greater representational flexibility. The first simulated dataset clearly shows that, when flexibility is lacking, the kernel mapping may not be able to fit the data, which in turn leads to poor performance for mappings that cannot capture the underlying structure, such as the linear or square kernels.

From the eigendecomposition perspective, the mixed kernel generally leads to larger r_{Corr} values than single kernels. Because this relationship is evident, we do not perform additional statistical tests.

Our observations partially align with previous research on mixed-kernel analysis (Zhu et al., 2012), with the difference that, in prior studies, "mixed" referred to combinations of bounded and unbounded kernels.

5.4. Computational cost

The computational costs for this study can be divided into two parts. One is from original fMRI to kernel space, which takes $O\left(T^2Q\right)$ time complexity. Back-reconstruction process, transferring the data from the kernel space to the original space, takes $O(TQ)$ time complexity. Even though the new back-reconstruction method is much more efficient in the second step, it does not totally avoid the time complexity generated by the mapping. Another difficulty is that the total problem is relatively large, as it contains four datasets, two methods, and eight different kernels.

As a practical example, one subject from the HCP dataset has dimensions $7\times390\times91\times109\times91$. Here, the first dimension 7 represents the number of spatial filters, the second corresponds to the time length, and the last three dimensions denote the spatial coordinates in MNI space. Storing these data in float32 precision requires approximately 10 GB of memory when using seven steerable filters. A typical commercial GPU, such as the NVIDIA GeForce RTX 3090 with 24 GB of memory, therefore, usually requires more than one GPU card to enable parallel processing, which in turn requires compatible hardware to operate them.

Due to the high dimensionality of fMRI data, a two-step approach may be preferable. These methods usually incorporate information from previous outputs, thereby introducing less randomness. The advantage is that less parallelism is needed; for example, we show that with only one or two shuffling, decent accuracy can be achieved. The disadvantage is that multi-step methods require a relatively complex design. In practice, on our workstation with an Intel Xeon w9-3475X CPU and 128 GB of memory, two subjects can be processed concurrently, so the mixed hyperbolic tangent kernel using the first back-reconstruction method requires approximately 6 hours of effective computation per subject.

5.5. Hyperparameter used in this study and validation

In Table 10, we list all the hyperparameters used, and their validation is described below. Thresholds used for simulated data generation are displayed in Table 2. For Simulation 1, similar activation patterns are adopted as those in Yang et al. (2020). Based on our practical experience, there is considerable flexibility in the choice of simulated data.

Table 10. List of hyperparameters in nonlinear KCCA, with detailed explanation and validation shown in Section 5.5.

Name	Value
fMRI simulation	Listed on Table 2
Smoothing level	FWHM = 4 mm
Kernel mapping	Listed on Table 3
Activation (voxel shuffling)	Top 10%
Reduce false positive	FPR < 0.1
Shuffling method 1	$Q_1 = Q_{\text{Non}}$, $Q_2 = Q_+$
Shuffling method 2	$Q_1 = Q_2 = 0.5Q_+$
Gray matter	Probability > 0.5
Activation (plotting)	$p < 0.05$

The threshold appearing in Method 3.2 specifies FWHM = 4 mm with seven steerable filters. Except for the 1D simulation in Section 2.7, all fMRI data, including both task-based and simulated data, use a 4 mm smoothing filter. The same smoothing level and steerable filters are reported in Yang et al. (2018), and a similar smoothing level is used in Cordes et al. (2012).

For nonlinear kernel parameters shown in Section 3.4, since both the linear and square kernels are normalized to have unit standard deviation, the mapping magnitude is chosen to match this scale. The selected range of γ covers a relatively large region. As stated in Section 3.7.2, optimization stability was assessed empirically by extending the optimization steps by 50%; the resulting relative change in $\mathcal{L}$ was below 1%, much smaller than the typical difference between kernels (greater than 20%).

The threshold appearing in Method 3.8 is based on the assumption that activation occupies approximately 10% of the total voxels within the brain. A similar metric is used in Yang et al. (2020). Another related consideration is reducing false positives in activation analysis. While a commonly used threshold of FPR < 0.1 is adopted, this criterion indicates that we expect the activation to remain relatively small (Skudlarski et al., 1999). As a result, we choose this value as the shuffling threshold. Two further validations are provided. First, for the first simulated dataset, the activation proportion is not exactly 10%, but typically ranges from 9% to 11%. Nevertheless, the performance improvement is evident. Second, even when shuffling is performed using 10% of the voxels, performance evaluated using entirely different metrics—AUC with FPR < 0.1 or the total AUC—shows clear improvement across different datasets.

The two shuffling thresholds are based on the activation thresholds described above. As stated in Section 5.1.4, this choice is mainly empirical due to high computational cost.

The threshold for gray matter is set to 0.5. We find that when this threshold is perturbed by 10%, from 0.45 to 0.55, the relative change in the number of voxels is less than 3%. Therefore, we believe that this threshold selection does not have a major influence on the results.

One related issue is the use of $p < 0.05$ to determine activation for plotting. This value is chosen solely for visualization purposes and does not affect the accuracy.

5.6. Subject- and group-level activation pattern

This section compares our subject-level analysis with the previous group-level analysis. First, all parameter optimization in this study is performed at the subject level without data separation. Here, "overfitting" refers to the case where parameters learned from nonlinear mappings focus solely on producing artificially large correlation values, while failing to recover biologically meaningful activation patterns.

Second, because activation is computed only at the subject level, the averages reported in this study are subject-wise averages. This differs from previous group-level analyses, in which fMRI data from the same group are concatenated along the time domain and activation is then computed at the group level. Notably, although the nonlinear KCCA algorithm has the capacity to perform such group-level computations, concatenating all data would require a very large amount of memory.

Compared with traditional group-level analysis, hippocampal and medial temporal lobe (MTL) activation in MCI shows a non-monotonic pattern rather than a simple reduction relative to normal aging. Several studies report reduced MTL or hippocampal activation in MCI during encoding or retrieval, consistent with neurodegenerative dysfunction (Johnson et al., 2004; Machulda et al., 2003; Petrella et al., 2006). In contrast, other studies demonstrate paradoxical hyperactivation in early or less impaired MCI, particularly during associative encoding tasks, despite minimal volumetric loss (Celone et al., 2006; Dickerson et al., 2005; Hämäläinen et al., 2007).

By simply comparing hippocampal activation using the ROI defined in Section 3.8.4 and evaluating the t-test between normal and MCI groups, both groups are considered across two tasks and two contrasts, leading to an effective sample size of 32 cases per group. The hypothesis is that hippocampal activation in the normal group is larger than that in the MCI group. The p-value for the linear kernel is 0.120, while it is 0.204 for the mixed hyperbolic tangent kernel, with neither reaching statistical significance. Therefore, our conclusion is that, in this study, we did not identify statistically significant differences between groups. Note that these results are based on subject-wise averages rather than a traditional group-level analysis.

For the HCP dataset, a similar analysis could be applied to examine gender-related differences in activation (Schmidt et al., 2009). However, as our study primarily focuses on methodological development, with activation extracted at the subject level, even if such differences were present, we would not expect them to affect our overall conclusions.

5.7. Sign effect in activation analysis

A critical question is whether to retain the sign of α during activation analysis. In previous GLM studies (Friman et al., 2003), the sign is preserved to indicate that only signals positively correlated with the effective design signal are selected. Conversely, earlier local CCA studies (Zhuang et al., 2017) adjusted weights to ensure positive correlation are used, resulting in activation patterns that may include both positive and negative components. In a previous publication, both cases are considered (Han et al., 2025).

In this study, we proposed a self-supervised learning method. The method presented here is specific to sign statistics, but we note that the method could also be extended to $|\alpha|$. In that scenario, the top 10% of voxels with the largest $|\alpha|$ values could be assumed to be activated, yielding two decision boundaries in the positive and negative regions. Similar shuffling along these boundaries could then be performed.

5.8. Neurological implications

Our findings suggest several neuroscience-relevant implications. First, the activations identified using the proposed nonlinear kernel framework showed substantially greater spatial overlap with gray-matter regions than with the linear model. Because BOLD task activation is physiologically expected to arise predominantly from gray-matter tissue, this improved alignment indicates that the method enhances detection of biologically meaningful activation rather than noise-driven fluctuations. Second, the nonlinear kernels, particularly the mixed hyperbolic tangent kernel, appear better able to model subject-specific variability and nonlinear relationships between neural responses and task conditions. This may be especially important in populations where functional organization is altered or less stereotyped across individuals. Finally, when examining canonical task-related regions of interest, the nonlinear model yielded activation patterns that were more localized to expected functional networks than the linear approach. Together, these results suggest that the proposed method may provide a more neurobiologically faithful representation of task-evoked functional activation and improve sensitivity to subtle but meaningful neural responses.

6. CONCLUSION

In this study, we extend the nonlinear KCCA method from previous work with several key contributions. First, we propose a self-supervised learning method that extracts kernel mappings purely by maximizing shuffling robustness. This method is more efficient, as it does not rely on additional datasets such as resting-state fMRI data, and it provides an alternative way to evaluate performance in the absence of ground truth. Second, we introduce a new back-reconstruction method, which is faster and more accurate than the previous version. In addition, the new method is more robust, with fewer subjects performing below the baseline. Third, we further investigate the kernel selection problem. By evaluating accuracy in fMRI activation detection, augmentation robustness, and eigendecomposition, we show that bounded kernels combining both linear and nonlinear terms outperform other kernels. Empirically, this improvement is attributed to the high noise self-correlation present in traditional kernel methods.

DATA AND CODE AVAILABILITY

The HCP data used in this study are publicly available and can be downloaded through the Human Connectome Project repository. Data access can be requested at https://www.humanconnectome.org/study/hcp-young-adult/data-releases. All the codes used in this study will be publicly available after acceptance at https://github.com/CCLRCBH-BIC.

AUTHOR CONTRIBUTIONS

Conceptualization: D.C.; methodology: C.H., Z.Y., X.Z., D.C.; supervision: D.C.; writing—initial draft: C.H.; writing—review and editing: C.H., Z.Y., X.Z., D.C.

DECLARATION OF COMPETING INTEREST

The authors have no conflict of interests to disclose.

ACKNOWLEDGMENTS

This study was supported by the National Institute of Aging (R01AG071566) and the National Institute of General Medical Sciences (P20GM109025).

REFERENCES

Alam, M. A. (2014). *Kernel choice for unsupervised kernel methods*. Graduate University for Advanced Studies, SOKENDAI. https://www.researchgate.net/profile/Md-Ashad-Alam/publication/297731182_Kernel_Choice_for_Unsupervised_Kernel_Methods/links/5b17faf7a6fdcca67b5d9b20/Kernel-Choice-for-Unsupervised-Kernel-Methods.pdf

Ashburner, J. (2009). Computational anatomy with the SPM software. *Magnetic Resonance Imaging*, *27*(8), 1163–1174. https://doi.org/10.1016/j.mri.2009.01.006

Bedel, H. A., Sivgin, I., Dalmaz, O., Dar, S. U., & Çukur, T. (2023). Bolt: Fused window transformers for fMRI time series analysis. *Medical Image Analysis*, *88*, 102841. https://doi.org/10.1016/j.media.2023.102841

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. https://doi.org/10.1016/j.neuroimage.2007.04.042

Bengio, Y., Delalleau, O., Roux, N. L., Paiement, J.-F., Vincent, P., & Ouimet, M. (2004). Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Computation*, *16*(10), 2197–2219. https://doi.org/10.1162/0899766041732396

Brodoehl, S., Gaser, C., Dahnke, R., Witte, O. W., & Klingner, C. M. (2020). Surface-based analysis increases the specificity of cortical activation patterns and connectivity results. *Scientific Reports*, *10*(1), 5737. https://doi.org/10.1038/s41598-020-62832-z

Bullmore, E., Long, C., Suckling, J., Fadili, J., Calvert, G., Zelaya, F., Carpenter, T. A., & Brammer, M. (2001). Colored noise and computational inference in neurophysiological (fMRI) time series analysis: Resampling methods in time and wavelet domains. *Human Brain Mapping*, *12*(2), 61–78. https://doi.org/10.1002/1097-0193(200102)12:2<61::aid-hbm1004>3.0.co;2-w

Celone, K. A., Calhoun, V. D., Dickerson, B. C., Atri, A., Chua, E. F., Miller, S. L., DePeau, K., Rentz, D. M., Selkoe, D. J., Blacker, D., Albert, M. S., & Sperling, R. A. (2006). Alterations in memory networks in mild cognitive impairment and Alzheimer's disease: An independent component analysis. *Journal of Neuroscience*, *26*(40), 10222–10231. https://doi.org/10.1523/jneurosci.2250-06.2006

Chen, B., Yang, J., Jeon, B., & Zhang, X. (2017). Kernel quaternion principal component analysis and its application in RGB-D object recognition. *Neurocomputing*, *266*, 293–303. https://doi.org/10.1016/j.neucom.2017.05.047

Chen, J. E., Jahanian, H., & Glover, G. H. (2017). Nuisance regression of high-frequency functional magnetic resonance imaging data: Denoising can be noisy. *Brain Connectivity*, *7*(1), 13–24. https://doi.org/10.1089/brain.2016.0441

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning* (pp. 1597–1607). https://doi.org/10.48550/arXiv.2002.05709

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning* (pp. 1310–1320). https://doi.org/10.48550/arXiv.1902.02918

Cole, M. W., Yarkoni, T., Repovš, G., Anticevic, A., & Braver, T. S. (2012). Global connectivity of prefrontal cortex predicts cognitive control and intelligence. *Journal of Neuroscience*, *32*(26), 8988–8999. https://doi.org/10.1523/jneurosci.0536-12.2012

Cordes, D., Jin, M., Curran, T., & Nandy, R. (2012). Optimizing the performance of local canonical correlation analysis in fMRI using spatial constraints. *Human Brain Mapping*, *33*(11), 2611–2626. https://doi.org/10.1002/hbm.21388

Cordes, D., & Nandy, R. (2007). Independent component analysis in the presence of noise in fMRI. *Magnetic Resonance Imaging*, *25*(9), 1237–1248. https://doi.org/10.1016/j.mri.2007.03.021

Dickerson, B., Salat, D., Greve, D., Chua, E., Rand-Giovannetti, E., Rentz, D., Bertram, L., Mullin, K., Tanzi,

R., Blacker, D., Albert, M. S., & Sperling, R. A. (2005). Increased hippocampal activation in mild cognitive impairment compared to normal aging and AD. *Neurology*, *65*(3), 404–411. https://doi.org/10.1212/01.wnl.0000171450.97464.49

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. https://doi.org/10.48550/arXiv.2010.11929

Erhardt, E. B., Allen, E. A., Wei, Y., Eichele, T., & Calhoun, V. D. (2012). SimTB, a simulation toolbox for fMRI data under a model of spatiotemporal separability. *NeuroImage*, *59*(4), 4160–4167. https://doi.org/10.1016/j.neuroimage.2011.11.088

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, j., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Fasshauer, G. E. (2007). *Meshfree approximation methods with matlab*. World Scientific Pub Co Inc. https://doi.org/10.1142/6437

Frackowiak, R. S., Friston, K. J., Frith, C. D., Dolan, R. J., Price, C. J., Zeki, S., Ashburner, J. T., & Penny, W. D. (2004). Chapter 40—analysis of fMRI time series: Linear time-invariant models, event-related fMRI, and optimal experimental design (2nd ed., 793–822). https://doi.org/10.1016/B978-012264841-0/50042-1

Friman, O., Borga, M., Lundberg, P., & Knutsson, H. (2003). Adaptive analysis of fMRI data. *NeuroImage*, *19*(3), 837–845. https://doi.org/10.1016/s1053-8119(03)00077-6

Friman, O., Cedefamn, J., Lundberg, P., Borga, M., & Knutsson, H. (2001). Detection of neural activity in functional mri using canonical correlation analysis. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *45*(2), 323–330. https://doi.org/10.1002/1522-2594(200102)45:2<323::aid-mrm1041>3.0.co;2-#

Gawryluk, J. R., Mazerolle, E. L., & D'Arcy, R. C. (2014). Does functional MRI detect activation in white matter? A review of emerging evidence, issues, and future directions. *Frontiers in Neuroscience*, *8*, 239. https://doi.org/10.3389/fnins.2014.00239

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., Jenkinson, M., & WU-Minn HCP Consortium. (2013). The minimal preprocessing pipelines for the human connectome project. *NeuroImage*, *80*, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, *9*(4), 416–429. https://doi.org/10.1006/nimg.1998.0419

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*. https://doi.org/10.1063/pt.5.028530

Gutmann, H.-M. (2001). A radial basis function method for global optimization. *Journal of Global Optimization*, *19*(3), 201–227. https://doi.org/10.1023/a:1011255519438

Hämäläinen, A., Pihlajamäki, M., Tanila, H., Hänninen, T., Niskanen, E., Tervo, S., Karjalainen, P. A., Vanninen, R. L., & Soininen, H. (2007). Increased fMRI responses during encoding in mild cognitive impairment. *Neurobiology of Aging*, *28*(12), 1889–1903. https://doi.org/10.1016/j.neurobiolaging.2006.08.008

Han, C., Yang, Z., Zhuang, X., & Cordes, D. (2024). Nonlinear kernel-based fMRI activation detection. In *Proceedings of the International Society for Magnetic Resonance in Medicine* (ISMRM). https://archive.ismrm.org/2024/3442.html

Han, C., Yang, Z., Zhuang, X., & Cordes, D. (2025). Nonlinear kernel-based fMRI activation detection. *Frontiers in Neuroimaging*, *4*, 1649749. https://doi.org/10.3389/fnimg.2025.1649749

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, *16*(12), 2639–2664. https://doi.org/10.1162/0899766042321814

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). IEEE. https://doi.org/10.1109/cvpr.2016.90

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708). IEEE. https://openaccess.thecvf.com/content_cvpr_2017/papers/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.pdf

Huang, S.-Y., Yeh, Y.-R., & Eguchi, S. (2009). Robust kernel principal component analysis. *Neural Computation*, *21*(11), 3179–3213. https://doi.org/10.1162/neco.2009.02-08-706

Jin, M., Pelak, V. S., Curran, T., Nandy, R. R., & Cordes, D. (2012). A preliminary study of functional abnormalities in AMCI subjects during different episodic memory tasks. *Magnetic Resonance Imaging*, *30*(4), 459–470. https://doi.org/10.1016/j.mri.2011.12.014

Johnson, S. C., Baxter, L. C., Susskind-Wilder, L., Connor, D. J., Sabbagh, M. N., & Caselli, R. J. (2004). Hippocampal adaptation to face repetition in healthy elderly and mild cognitive impairment. *Neuropsychologia*, *42*(7), 980–989. https://doi.org/10.1016/j.neuropsychologia.2003.11.015

Karimi, D. (2024). An approach to building foundation models for brain image analysis. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, *15012*, 421–431. https://doi.org/10.1007/978-3-031-72390-2_40

Kim, B.-H., Ye, J. C., & Kim, J.-J. (2021). Learning dynamic graph representation of brain connectome with spatio-temporal attention. *Advances in Neural Information Processing Systems*, *34*, 4314–4327. https://doi.org/10.52202/068431-0440

Kim, P., Kwon, J., Joo, S., Bae, S., Lee, D., Jung, Y., Yoo, S., Cha, J., & Moon, T. (2023). SwiFT: Swin 4D fMRI transformer. *Advances in Neural Information Processing Systems*, *36*, 42015–42037. https://doi.org/10.52202/075280-1820

Kwon, J., Seo, J., Wang, H., Moon, T., Yoo, S., & Cha, J. (2025). Predicting task-related brain activity from resting-state brain dynamics with fMRI transformer. *Imaging Neuroscience*, *3*, imag_a_00440. https://doi.org/10.1162/imag_a_00440

Laird, A. R., Rogers, B. P., & Meyerand, M. E. (2004). Comparison of Fourier and wavelet resampling methods. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *51*(2), 418–422. https://doi.org/10.1002/mrm.10671

Lamichhane, B., Westbrook, A., Cole, M. W., & Braver, T. S. (2020). Exploring brain-behavior relationships in the n-back task. *NeuroImage*, *212*, 116683. https://doi.org/10.1016/j.neuroimage.2020.116683

Langs, G., Menze, B. H., Lashkari, D., & Golland, P. (2011). Detecting stable distributed patterns of brain activation using Gini contrast. *NeuroImage*, *56*(2), 497–507. https://doi.org/10.1016/j.neuroimage.2010.07.074

Li, J., Chen, J., Tang, Y., Wang, C., Landman, B. A., & Zhou, S. K. (2023). Transforming medical imaging with transformers? A comparative review of key properties, current progresses, and future perspectives. *Medical Image Analysis*, *85*, 102762. https://doi.org/10.1016/j.media.2023.102762

Li, X. (2014). *Functional magnetic resonance imaging processing*. *Springer*. https://doi.org/10.1007/978-94-007-7302-8

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. https://doi.org/10.1016/j.neuroimage.2008.10.065

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022). IEEE. https://doi.org/10.1109/iccv48922.2021.00986

Liu, Z., Liu, Q., Liu, T., Xu, N., Lin, X., Wang, Y., & Wen, W. (2019). Feature distillation: DNN-oriented JPEG compression against adversarial examples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 860–868). IEEE. https://doi.org/10.1109/cvpr.2019.00095

Logothetis, N. K., & Wandell, B. A. (2004). Interpreting the bold signal. *Annual Review of Physiology*, *66*, 735–769. https://doi.org/10.1146/annurev.physiol.66.082602.092845

Machulda, M., Ward, H., Borowski, B., Gunter, J., Cha, R., O'brien, P., Petersen, R., Boeve, B., Knopman, D., Tang-Wai, D., Ivnik, R. J., Smith, G. E., Tangalos, E. G., & Jack , Jr, C. R. (2003). Comparison of memory fMRI response among normal, MCI, and Alzheimer's patients. *Neurology*, *61*(4), 500–506. https://doi.org/10.1212/01.wnl.0000079052.01016.78

Mejia, A. F., Yue, Y. R., Bolin, D., Lindgren, F., & Lindquist, M. A. (2020). A Bayesian general linear modeling approach to cortical surface fMRI data analysis. *Journal of the American Statistical Association*, *115*(530), 501–520. https://doi.org/10.1080/01621459.2019.1611582

Mo, S., Kang, H., Sohn, K., Li, C.-L., & Shin, J. (2021). Object-aware contrastive learning for debiased scene representation. *Advances in Neural Information Processing Systems*, *34*, 12251–12264. https://doi.org/10.52202/068431-0468

Nandy, R. R., & Cordes, D. (2003). Novel ROC-type method for testing the efficiency of multivariate statistical methods in fMRI. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *49*(6), 1152–1162. https://doi.org/10.1002/mrm.10469

Nandy, R. R., & Cordes, D. (2004). New approaches to receiver operating characteristic methods in functional magnetic resonance imaging with real data using repeated trials. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *52*(6), 1424–1431. https://doi.org/10.1002/mrm.20263

Petrella, J. R., Krishnan, S., Slavin, M. J., Tran, T.-T. T., Murty, L., & Doraiswamy, P. M. (2006). Mild cognitive impairment: Evaluation with 4-T functional MR imaging. *Radiology*, *240*(1), 177–186. https://doi.org/10.1148/radiol.2401050739

Prakash, A., Moran, N., Garber, S., DiLillo, A., & Storer, J. (2018). Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8571–8580). IEEE. https://doi.org/10.1109/cvpr.2018.00894

Rasmussen, P. M., Madsen, K. H., Lund, T. E., & Hansen, L. K. (2011). Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. *NeuroImage*, *55*(3), 1120–1131. https://doi.org/10.1016/j.neuroimage.2010.12.035

Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A. R., Schulz, J. B., Fox, P. T., & Eickhoff, S. B. (2012). Modelling neural correlates of working memory: A coordinate-based meta-analysis. *NeuroImage*, *60*(1), 830–846. https://doi.org/10.1016/j.neuroimage.2011.11.050

Salazar, J. Z., Kwakkel, J. H., & Witvliet, M. (2024). Evaluating the choice of radial basis functions in multiobjective optimal control applications. *Environmental Modelling & Software*, *171*, 105889. https://doi.org/10.1016/j.envsoft.2023.105889

Schilling, K. G., Li, M., Rheault, F., Gao, Y., Cai, L., Zhao, Y., Xu, L., Ding, Z., Anderson, A. W., Landman, B. A., & Gore, J. C. (2023). Whole-brain, gray, and white matter time-locked functional signal changes with simple tasks and model-free analysis. *Proceedings of the National Academy of Sciences*, *120*(42), e2219666120. https://doi.org/10.1073/pnas.2219666120

Schmidt, H., Jogia, J., Fast, K., Christodoulou, T., Haldane, M., Kumari, V., & Frangou, S. (2009). No gender differences in brain activation during the n-back task: An fMRI study in healthy individuals. *Human Brain Mapping*, *30*(11), 3609–3615. https://doi.org/10.1002/hbm.20783

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626). IEEE. https://doi.org/10.1109/iccv.2017.74

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press. https://doi.org/10.1017/cbo9780511809682

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*. https://doi.org/10.48550/arXiv.1312.6034

Skudlarski, P., Constable, R. T., & Gore, J. C. (1999). Roc analysis of statistical methods used in functional MRI: Individual subjects. *NeuroImage*, *9*(3), 311–329. https://doi.org/10.1006/nimg.1999.0402

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). SmoothGrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825*. https://doi.org/10.48550/arXiv.1706.03825

Sorenson, J. A., & Wang, X. (1996). ROC methods for evaluation of fMRI techniques. *Magnetic Resonance in Medicine*, *36*(5), 737–744. https://doi.org/10.1002/mrm.1910360512

Tamkin, A., Glasgow, M., He, X., & Goodman, N. (2023). Feature dropout: Revisiting the role of augmentations in contrastive learning. *Advances in Neural Information Processing Systems*, *36*, 61806–61835. https://doi.org/10.52202/075280-2702

Tang, Y., Yang, D., Li, W., Roth, H. R., Landman, B., Xu, D., Nath, V., & Hatamizadeh, A. (2022). Self-supervised pre-training of swin transformers for 3D medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20730–20740). IEEE. https://doi.org/10.1109/cvpr52688.2022.02007

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot,

M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289. https://doi.org/10.1006/nimg.2001.0978

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., Penna, S. D., Feinberg, D., Glasser, M. F., Harel, N., Heath, A. C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., … WU-Minn HCP Consortium. (2012). The human connectome project: A data acquisition perspective. *NeuroImage*, *62*(4), 2222–2231. https://doi.org/10.1016/j.neuroimage.2012.02.018

Wang, Z. (2009). A hybrid SVM–GLM approach for fMRI data analysis. *NeuroImage*, *46*(3), 608–615. https://doi.org/10.1016/j.neuroimage.2009.03.0ss16

Wang, Z., Childress, A. R., Wang, J., & Detre, J. A. (2007). Support vector machine learning-based fMRI data group analysis. *NeuroImage*, *36*(4), 1139–1151. https://doi.org/10.1016/j.neuroimage.2007.03.072

Xifra-Porxas, A., Ghosh, A., Mitsis, G. D., & Boudrias, M.-H. (2021). Estimating brain age from structural MRI and MEG data: Insights from dimensionality reduction techniques. *NeuroImage*, *231*, 117822. https://doi.org/10.1016/j.neuroimage.2021.117822

Yakupov, R., Lei, J., Hoffmann, M. B., & Speck, O. (2017). False fMRI activation after motion correction. *Human Brain Mapping*, *38*(9), 4497–4510. https://doi.org/10.1002/hbm.23677

Yang, Z., Zhuang, X., Lowe, M., & Cordes, D. (2022). A global spatially-adaptive method for task fMRI activation analysis with the advantage of alleviated spatial blurring. In *Proceedings of the International Society for Magnetic Resonance in Medicine (ISMRM)*. https://archive.ismrm.org/2022/3340.html

Yang, Z., Zhuang, X., Lowe, M. J., & Cordes, D. (2025). A deep neural network for adaptive spatial smoothing of task fMRI data. *Frontiers in Neuroimaging*, *4*, 1554769. https://doi.org/10.3389/fnimg.2025.1554769

Yang, Z., Zhuang, X., Sreenivasan, K., Mishra, V., Curran, T., Byrd, R., Nandy, R., & Cordes, D. (2018). 3D spatially-adaptive canonical correlation analysis: Local and global methods. *NeuroImage*, *169*, 240–255. https://doi.org/10.1016/j.neuroimage.2017.12.025

Yang, Z., Zhuang, X., Sreenivasan, K., Mishra, V., Curran, T., & Cordes, D. (2020). A robust deep neural network for denoising task-based fMRI data: An application to working memory and episodic memory. *Medical Image Analysis*, *60*, 101622. https://doi.org/10.1016/j.media.2019.101622

Yu, X., Zhang, L., Zhao, L., Lyu, Y., Liu, T., & Zhu, D. (2022). Disentangling spatial-temporal functional brain networks via twin-transformers. *arXiv:2204.09225*. https://doi.org/10.48550/arXiv.2204.09225

Zhang, J., & Ma, K. (2022). Rethinking the augmentation module in contrastive learning: Learning hierarchical augmentation invariance with expanded views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16650–16659). IEEE. https://doi.org/10.1109/cvpr52688.2022.01615

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921–2929). IEEE. https://doi.org/10.1109/cvpr.2016.319

Zhou, Z., Sodha, V., Pang, J., Gotway, M. B., & Liang, J. (2021). Models genesis. *Medical Image Analysis*, *67*, 101840. https://doi.org/10.1016/j.media.2020.101840

Zhu, X., Huang, Z., Shen, H. T., Cheng, J., & Xu, C. (2012). Dimensionality reduction by mixed kernel canonical correlation analysis. *Pattern Recognition*, *45*(8), 3003–3016. https://doi.org/10.1016/j.patcog.2012.02.007

Zhuang, X., Yang, Z., & Cordes, D. (2020). A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping*, *41*(13), 3807–3833. https://doi.org/10.1002/hbm.25090

Zhuang, X., Yang, Z., Curran, T., Byrd, R., Nandy, R., & Cordes, D. (2017). A family of locally constrained CCA models for detecting activation patterns in fMRI. *NeuroImage*, *149*, 63–84. https://doi.org/10.1016/j.neuroimage.2016.12.081

APPENDIX A

A.1. Steerable filters

Suppose F_{orig} is the Gaussian filter with FWHM = 4 mm used in GLM, and G_{iso} is defined as a Gaussian function with FWHM equal to half of F_{orig} (FWHM = 2 mm). Let G_i ($i = 1,\ldots,6$) denote

$$
\begin{aligned}
&G_i(\mathbf{x}) = \left(1 - G_{\text{iso}}(\mathbf{x})\right)\left[\left(\hat{\mathbf{n}}_i^T \hat{\mathbf{x}}\right) - \frac{1}{6}\right],\\
&\hat{\mathbf{n}}_1 = \begin{pmatrix} a \\ 0 \\ b \end{pmatrix}, \hat{\mathbf{n}}_2 = \begin{pmatrix} -a \\ 0 \\ b \end{pmatrix}, \hat{\mathbf{n}}_3 = \begin{pmatrix} b \\ a \\ 0 \end{pmatrix},\\
&\hat{\mathbf{n}}_4 = \begin{pmatrix} a \\ 0 \\ b \end{pmatrix}, \hat{\mathbf{n}}_5 = \begin{pmatrix} -a \\ 0 \\ b \end{pmatrix}, \hat{\mathbf{n}}_6 = \begin{pmatrix} b \\ a \\ 0 \end{pmatrix},\\
&a = \frac{2}{\sqrt{10 + 2\sqrt{5}}}, b = \frac{1 + \sqrt{5}}{\sqrt{10 + 2\sqrt{5}}}.
\end{aligned}
\tag{A1}
$$

Seven steerable filters, F_{iso} and F_i ($i = 1,\ldots,6$), are defined as

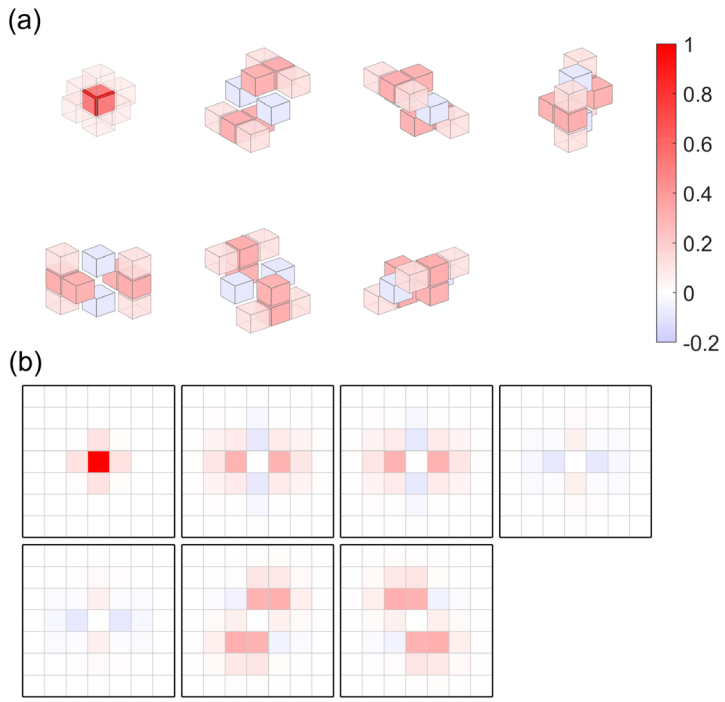
$$
F_{\text{iso}}(\mathbf{x}) = G_{\text{iso}}(\mathbf{x}), F_i(\mathbf{x}) = G_i(\mathbf{x}) F_{\text{orig}}(\mathbf{x}) \text{ for } i = 1,\cdots,6. \tag{A2}
$$

We can transform those filters to a matrix form $\mathbf{A}$ and decompose it to $\mathbf{A}^{(0,\cdots,6)}$. After expanding the smooth kernel to a $7\times7\times7$ grid, we plotted the result in Appendix Figure A1.

A.2. Normalization for the kernel method

In Table 3 all the kernels can be generated using the linear or square kernel. We introduce two normalization constants

$$
\begin{aligned}
\mathcal{N}_{\text{Linear}} &= \frac{1}{T^2}\sum_i\sum_j\left|\sum_p\left(\tilde{Y}_{ip}\tilde{Y}_{jp}\right)\right|,\\
\mathcal{N}_{\text{Square}} &= \frac{1}{T^2}\sum_i\sum_j\left|\sum_p\left(\tilde{Y}_{ip} - \tilde{Y}_{ip}\right)^2\right|.
\end{aligned}
\tag{A3}
$$



Appendix Fig. A1. (a) Seven $7\times7\times7$ steerable filters are used for 3D fMRI data with grid size 2 mm. The equal weight sum is equal to a single Gaussian smoothing filter with FWHM = 4 mm. (b) The same as (a) but projected on the x-axis.

To simplify notation, we substitute $K_{\mathbf{Y}}$ in the main article by $K_{(\text{Kernel name})}$ as there is no ambiguity in this context. Together with the normalization, eight kernels are defined as

$$
\begin{aligned}
\left(K_{\text{Linear}}\right)_{ij} &= \frac{1}{\mathcal{N}_{\text{Linear}}}\sum\nolimits_p \widetilde{Y}_{ip}\widetilde{Y}_{jp},\\
\left(K_{\text{Parabolic}}\right)_{ij} &= \left(\frac{1}{\mathcal{N}_{\text{Linear}}}\sum\nolimits_p \widetilde{Y}_{ip}\widetilde{Y}_{jp}+b^2\right)^2,\\
\left(K_{\text{Gaussian}}\right)_{ij} &= \exp\left[-\frac{1}{\sigma^2}\frac{1}{\mathcal{N}_{\text{Square}}}\sum\nolimits_p\left(\widetilde{Y}_{ip}-\widetilde{Y}_{jp}\right)^2\right],\\
\left(K_{\text{Inverse}}\right)_{ij} &= 1/\sqrt{\sum\nolimits_p \frac{1}{\mathcal{N}_{\text{Square}}}\left(\widetilde{Y}_{ip}-\widetilde{Y}_{jp}\right)^2+b^2},\\
\left(K_{\text{Bounded Linear}}\right)_{ij} &= \begin{cases}\left(K_{\text{Linear}}\right)_{ij} & \text{For } \left(K_{\text{Linear}}\right)_{ij} <= C,\\ C & \text{For } \left(K_{\text{Linear}}\right)_{ij} > C.\end{cases},\\
\left(K_{\text{Square}}\right)_{ij} &= \frac{1}{\mathcal{N}_{\text{Square}}}\sum\nolimits_p\left(\widetilde{Y}_{ip}-\widetilde{Y}_{jp}\right)^2,\\
\left(K_{\text{Tanh}}\right)_{ij} &= \tanh\left(b\sum\nolimits_p \frac{1}{\mathcal{N}_{\text{Linear}}}\widetilde{Y}_{ip}\widetilde{Y}_{jp}+c\right),\\
\left(K_{\text{Mixed Tanh}}\right)_{ij} &= \tanh\left[b_1\frac{1}{\mathcal{N}_{\text{Linear}}}\sum\nolimits_p \widetilde{Y}_{ip}\widetilde{Y}_{jp}+b_2\frac{1}{\mathcal{N}_{\text{Square}}}\right.\\
&\quad \left.\sum\nolimits_p\left(\widetilde{Y}_{ip}-\widetilde{Y}_{jp}\right)^2+c\right].
\end{aligned}
\tag{A4}
$$

A.3. Proof of Eq. (5)

In Eq. (4), we insert the identity matrix $I=\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T\left(\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T\right)^{\dagger}$ on the right.

$$
\begin{aligned}
&\mathbf{Y}\alpha-\left(\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T\left(\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T\right)^{\dagger}\right)K_{\mathbf{Y}}\mathbf{v}_{\mathbf{Y}}=0\\
&\mathbf{Y}\left[\alpha-\mathbf{A}\widetilde{\mathbf{Y}}^T\left(\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T\right)^{\dagger}K_{\mathbf{Y}}\mathbf{v}_{\mathbf{Y}}\right]=0.
\end{aligned}
\tag{A5}
$$

Let the bracket be equal to 0, then we have

$$
\alpha=\mathbf{A}\widetilde{\mathbf{Y}}^T\left(\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T\right)^{\dagger}K_{\mathbf{Y}}\mathbf{v}_{\mathbf{Y}}. \tag{A6}
$$

One practical difficulty is that the eigenvector $\mathbf{v}_{\mathbf{Y}}$ contains the uncertainty for the sign. Note that the $\mathbf{X}_{\text{eff}}$ is represented by the kernel space term $K_{\mathbf{X}}\mathbf{v}_{\mathbf{X}}$, then in the kernel space the relationship between $K_{\mathbf{X}}\mathbf{v}_{\mathbf{X}}$ and $K_{\mathbf{Y}}\mathbf{v}_{\mathbf{Y}}$ is maximized. To make the sign consistent, we introduce $s=\text{sign}\left[\text{corr}\left(\mathbf{X}_{\text{eff}},K_{\mathbf{Y}}\mathbf{v}_{\mathbf{Y}}\right)\right]$. Finally,

$$
\alpha=s\mathbf{A}\widetilde{\mathbf{Y}}^T\left(\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T\right)^{\dagger}K_{\mathbf{Y}}\mathbf{v}_{\mathbf{Y}}. \tag{A7}
$$

As the maximum rank for $\mathbf{Y}$ is T, removing $\mathbf{Y}$ outside the bracket introduces other constraints for the solution. From the relationship between the unknown and the number of equations, the original Eq. (4) only has T constraints. By removing $\mathbf{Y}$ we increase the constraints to Q.

A.4. Proof of Eq. (6)

Starting from Eq. (4), both sides are represented by a vector with length t. For any t we have

$$\mathbf{Y}_t\alpha - \left(K_{\mathbf{Y}}\mathbf{v}_{\mathbf{Y}}\right)_t = 0. \tag{A8}$$

Taken the derivative for Y_t on both sides yields

$$\alpha = \frac{\partial\left(K_{\mathbf{Y}}\mathbf{v}_{\mathbf{Y}}\right)_t}{\partial\mathbf{Y}_t}. \tag{A9}$$

Practically evaluating this equation has two difficulties. α should satisfy this equation for all times t, and $\mathbf{v}_{\mathbf{Y}}$ is the eigenvector corresponding to the largest eigenvalues and it is not differentiable. We take an approximation for this problem by using summation from different times and treat $\mathbf{v}_{\mathbf{Y}}$ as a constant during differentiation. Similar to the analyses shown in Section 7.2, we add the sign to remove the uncertainty so that

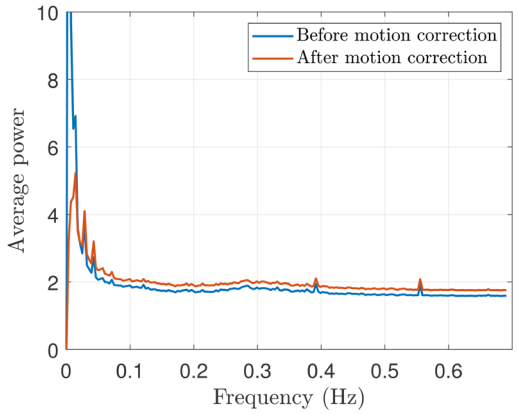
$$\alpha = s\sum_{t_1t_2}\frac{\partial\left(K_{\mathbf{Y}}\right)_{t_1t_2}}{\partial\mathbf{Y}_{t_1}}\left(\mathbf{v}_{\mathbf{Y}}\right)_{t_2}. \tag{A10}$$

From the relationship between the unknown and the number of equations, the original Eq. (4) only has T constraints. By taking derivative, we increased the number of constraints to $Q\times T$. Note that Eq. (A9) is defined for each time t. Finally we reduce the constraints to Q by taking an average of all time points.

A.5. Proof of equivalence for the linear kernel

For BC 1, we insert $K_{\mathbf{Y}} = \widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T$ so that

$$\alpha = s\mathbf{A}\widetilde{\mathbf{Y}}^T\left(\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T\right)^{\dagger}\left(\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T\right)\mathbf{v}_{\mathbf{Y}} = s\mathbf{A}\mathbf{A}^T\mathbf{Y}^T\mathbf{v}_{\mathbf{Y}}. \tag{A11}$$



Appendix Fig. A2. Average power spectral density versus frequency before and after motion correction. Results are averaged across 87 subjects from the HCP dataset.

For BC 2, we use the equation from Han et al. (2025) and obtain

$$\alpha = \mathbf{A}\tilde{\alpha}, \text{ and } \tilde{\alpha}_p = \sum_{ij}\frac{\partial\left(K_{\mathbf{Y}}\right)_{ij}}{\partial\tilde{Y}_{ip}}\left(\mathbf{v}_{\mathbf{Y}}\right)_j. \tag{A12}$$

Then,

$$\begin{aligned}\tilde{\alpha}_p &= \sum_{ij}\frac{\partial\left(K_{\mathbf{Y}}\right)_{ij}}{\partial\tilde{Y}_{ip}}\left(\mathbf{v}_{\mathbf{Y}}\right)_j \\ &= \frac{1}{\mathcal{N}_{\text{linear}}}\sum_{ij}\frac{\partial\sum_x\tilde{Y}_{ix}\tilde{Y}_{jx}}{\partial\tilde{Y}_{ip}}\left(\mathbf{v}_{\mathbf{Y}}\right)_j \\ &= \frac{1}{\mathcal{N}_{\text{linear}}}\sum_{ij}\left[\tilde{Y}_{jp}\left(\mathbf{v}_{\mathbf{Y}}\right)_j + \delta_{ij}\tilde{Y}_{jp}\left(\mathbf{v}_{\mathbf{Y}}\right)_j\right] \\ &= \frac{T+1}{\mathcal{N}_{\text{linear}}}\tilde{\mathbf{Y}}_p^T\mathbf{v}.\end{aligned} \tag{A13}$$

Then, for BC 2 we obtain

$$\alpha = \frac{T+1}{\mathcal{N}_{\text{linear}}}\mathbf{A}\tilde{\mathbf{Y}}^T\mathbf{v}_{\mathbf{Y}} = \frac{T+1}{\mathcal{N}_{\text{linear}}}\mathbf{A}\mathbf{A}^T\mathbf{Y}^T\mathbf{v}_{\mathbf{Y}}. \tag{A14}$$

A.6. Power analysis before and after motion correction

Six motion parameters from the HCP dataset are denoted as $\mathbf{z}_{\text{motion}} \in \Re^{T\times 6}$ (Van Essen et al., 2012). A linear regression model is applied to remove motion-related components:

$$\mathbf{Y}_{\text{correction}} = \mathbf{Y} - \mathbf{z}_{\text{motion}}\mathbf{e}, \tag{A15}$$

where $\mathbf{e} = \left(\mathbf{z}_{\text{motion}}\right)^{\dagger}\mathbf{Y} \in \Re^{6\times Q}$.

Power spectral analysis based on the discrete Fourier transform is then performed on both $\mathbf{Y}$ and $\mathbf{Y}_{\text{correction}}$ for comparison. Prior to the analysis, each voxel time series is normalized to zero mean and unit variance in the time domain. The resulting Fourier power spectra are shown in Appendix Figure A2, averaged across all 87 subjects in the HCP dataset.

The results show that motion correction reduces power density in the low-frequency range, while increasing power at higher frequencies (>0.05 Hz). This suggests that including motion regressors may introduce additional high-frequency artifacts in the HCP dataset.