Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Anatomical Priors and Voxel Selection as Interchangeable Dimensionality Controls in fMRI Decoding: A Leakage-Free Study

Tanay Chowdhury

tanayz@outlook.com

Amazon (United States) https://orcid.org/0009-0001-5373-7458

Research Article

Keywords: functional MRI, multivariate pattern analysis, feature selection, ventral temporal cortex, cross-validation, data leakage

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10828698/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Anatomical Priors and Voxel Selection as Interchangeable Dimensionality Controls in fMRI Decoding: A Leakage-Free Study

Tanay Chowdhury
Independent Researcher
tanayz@outlook.com

Abstract

Decoding from functional MRI (fMRI) must reduce tens of thousands of voxels to a number a classifier can fit from tens to hundreds of examples, either with an anatomical prior (a mask of the region believed to carry the signal) or with data-driven voxel selection. We compared the two under nested, run-grouped cross-validation in which every selection and hyperparameter choice saw only training data, with permutation tests repeating the entire nested procedure. In the six-subject Haxby object-category dataset, univariate selection improved whole-brain decoding by +25.4 points (0.434 to 0.688) but added nothing within a ventral temporal (VT) mask: pooled over 9 subjects in two datasets, selection changed accuracy by −0.8 points, equivalent to zero within ±3 points. Label-blind reductions (random voxels, variance filter, principal components) did not reproduce the whole-brain gain. The two controls are interchangeable because they find the same cortex: without any spatial information, the ranking placed 15–33% of its voxels in VT (at least 12% in every fold), which holds 0.8–1.7% of the brain. VT was sufficient but not necessary: without it, the rest of the brain still decoded at 0.606 (chance 0.125), and in three subjects whole-brain selection outperformed the mask. On permuted labels, the ungrouped cross-validation of a widely used tutorial, since corrected, reported 21.2% (chance 12.5%), and ranking voxels before the split reported 92.5–95.2% (chance 50%). A closed-form approximation predicts the second bias from sample size and voxel counts. Maps of informative voxels are valid only when selection is nested inside the evaluation.

Keywords: functional MRI; multivariate pattern analysis; feature selection; ventral temporal cortex; cross-validation; data leakage

1 Introduction

Decoding analyses of functional MRI (fMRI) ask whether, and where, a pattern of activity carries information about a stimulus or a cognitive state (Haynes and Rees, 2006; Norman et al., 2006; Pereira et al., 2009). A volume contains on the order of 10^4 voxels, whereas a scanning session yields tens to hundreds of labeled examples, so the candidate pool must be reduced before a classifier can be fit. Two remedies are standard. An *anatomical prior* restricts the analysis in advance to a region believed to carry the signal. A *data-driven selection* step ranks voxels by a statistic computed from the labels and keeps the top k (Guyon and Elisseeff, 2003; De Martino et al., 2008; Chu et al., 2012). Both reduce the variance of a linear classifier at the risk of discarding signal, and the number of voxels that decodes best can vary considerably between individuals (Bleichner et al., 2014). Yet the two are rarely compared on the same data under the same evaluation, and whether a selection step is worth having is usually settled by convention. Chu et al. (2012) found that univariate selection did not improve whole-brain support vector machines on a large structural dataset, and De Martino

et al. (2008) combined univariate filtering with recursive elimination, but neither compared selection with an anatomical prior.

The Haxby object-category data (Haxby et al., 2001) make the contrast concrete. Ventral temporal (VT) cortex is the canonical locus of object-category information: category-selective regions for faces (Kanwisher et al., 1997), places (Epstein and Kanwisher, 1998) and objects in general (Malach et al., 1995) lie along the ventral occipitotemporal surface, and Haxby et al. (2001) showed that eight categories are represented there by distributed, overlapping response patterns. How far that information extends beyond VT has been debated (Spiridon and Kanwisher, 2002; O'Toole et al., 2005; Kriegeskorte et al., 2008; Grill-Spector and Weiner, 2014), and part of it may reflect low-level image properties that covary with category (Rice et al., 2014). The standard analysis of these data restricts the classifier to each subject's VT mask (Haxby et al., 2011, 2014), whereas the widely used `nilearn` tutorial on the same data (Abraham et al., 2014) selects voxels from the whole brain by a univariate statistic. These are two answers to the same statistical problem.

Comparing them requires an evaluation that cannot leak. Selection uses the labels, so ranking voxels on all the data before cross-validation biases the estimate upward (Ambroise and McLachlan, 2002; Cawley and Talbot, 2010), and the bias grows with the ratio of candidate voxels to examples; in neuroimaging this is known as "double dipping" (Kriegeskorte et al., 2009). Hyperparameters tuned on the evaluation folds bias the estimate in the same way, and nested cross-validation is the standard remedy (Cawley and Talbot, 2010; Hastie et al., 2009; Varoquaux et al., 2017). Time series add a second route: when temporally adjacent volumes share a label and correlated noise, folds that split them leak through autocorrelation (Varoquaux et al., 2017; Varoquaux, 2018). Both errors are instances of the leakage that Kaufman et al. (2012) formalized and that Kapoor and Narayanan (2023) found across many fields of machine-learning-based science. Both also corrupt a second, common use of selection: maps of where the "informative" voxels lie. Even without leakage, cross-validation estimates at typical fMRI sample sizes have errors of about ± 10 points (Varoquaux, 2018), and what a decodable pattern licenses about representation is limited (Hebart and Baker, 2018; Poldrack et al., 2020). Our contribution is therefore not a new form of leakage but a measurement, on real and simulated null data, of how large the two most common forms are in a setting where both occur, together with an ablation that is only interpretable once both are removed.

We compare the two dimensionality controls on two public fMRI datasets with a strictly nested, run-grouped protocol in which every selection and hyperparameter choice sees training data only, and with permutation tests that repeat the entire nested procedure. The selection step is a summed one-versus-rest F ranking that generalizes the active-voxel method of Chowdhury (2014) to multi-class problems. It uses no spatial information, which makes it an anatomy-blind probe of where a linear decoder finds its signal. Our results are:

1. **Selection helps only without a prior.** On the six-subject Haxby data, selection improves whole-brain decoding by +25.4 points, positive in every subject and not replaced by stronger ℓ_2 regularization, but is statistically equivalent to using every voxel once the VT mask is applied, on Haxby and on a second dataset with a different paradigm (StarPlus; Mitchell et al., 2004). The whole-brain gain is reproduced by a second ranking statistic and by linear and Gaussian-kernel support vector machines, but not by reductions that ignore the labels (random voxels, a variance filter, principal components): what matters is which voxels are kept, not how many. Within the prior, the equivalence extends to a linear and to a tuned Gaussian-kernel support vector machine (Sect. 3.1).

2. **The two controls select the same cortex.** Without any spatial information, the ranking concentrates in VT, 9–35 times more than chance on average and significantly so in every fold

of every subject; its other voxels lie mostly more than 10 mm from VT, and 42–80% of them in the posterior third of the brain. The prior is sufficient but not necessary: the remaining voxels still decode far above chance, and in half the subjects data-driven selection outperforms the prior (Sects. 3.2–3.3).

3. **What leaky protocols report on null data.** On permuted labels, the ungrouped cross-validation of the `nilearn` tutorial reports 21.2% accuracy for an eight-class problem (the tutorial has since been changed to run-grouped folds, following the preprint of this work; Nilearn contributors, 2026a,b), and selection before the split reports 92.5–95.2% for a two-class problem. A ground-truth simulation separates the two mechanisms: one scales with temporal autocorrelation, the other with dimensionality. For the second we give a closed-form approximation, Eq. (5), that predicts the accuracy reported on pure noise from the numbers of examples, candidate voxels and selected voxels, and matches simulation within 1.7 points (Sect. 3.4).

The practical conclusion, for univariate selection and the linear and kernel decoders studied here, is that a selection step and a defensible anatomical prior are substitutes, not complements, and that maps derived from selection are meaningful only when selection is nested inside the evaluation.

2 Materials and Methods

2.1 Datasets and candidate pools

Haxby. Six subjects viewed grayscale images of eight categories (bottles, cats, chairs, faces, houses, scissors, scrambled pictures, shoes) in a block design across 12 scanner runs (11 for subject 5), TR = 2.5 s (Haxby et al., 2001). We use `nilearn`'s loader (Abraham et al., 2014), apply 4 mm smoothing and per-run detrending within the whole-brain mask as the library's example does, and exclude rest, leaving 864 labeled volumes per subject (792 for subject 5). All analyses are within subject in native space ($40 \times 64 \times 64$ voxels of $3.5 \times 3.75 \times 3.75$ mm); the whole-brain mask contains 39,912 voxels. The subject-specific VT masks distributed with the data (307–675 voxels, 0.8–1.7% of the brain) are the anatomical prior; VT is the region in which the original study located object-category information (Haxby et al., 2001), and the standard analysis of these data decodes from it (Haxby et al., 2011, 2014). The face- and house-selective masks distributed with the data are used only to characterize selected voxels. Subject 6 has no anatomical image, so maps are shown on each subject's mean functional image.

StarPlus. Three subjects (04799, 04820, 04847) performed a picture-versus-sentence verification task, sampled every 500 ms (Mitchell et al., 2004). Following the `fmri_core` toolbox distributed with the data, each of 40 trials per subject yields one picture and one sentence example (80 per subject, 40 per class), and the two examples of a trial are always kept in the same fold. Features are single-trial beta-series estimates (Rissman et al., 2004): a first-level linear model with one regressor per event (least-squares-all, canonical hemodynamic response) whose effect maps are the patterns (∼4,700–5,000 voxels in the distributed regions of interest).

2.2 Selection score and classifier

For one subject let $\mathcal{D} = \{(\mathbf{x}_i, y_i, g_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^p$ holds the candidate voxels, $y_i \in \{1, \ldots, M\}$ the class ($M = 8$ for Haxby, 2 for StarPlus) and g_i the group that must never be split between training and test (scanner run for Haxby, trial for StarPlus). For an index set T, $\mathcal{D}_T$ denotes those

Algorithm 1 Nested, group-respecting evaluation (one subject)

Require: $\mathcal{D}$, groups g, grid Θ with $k \in \mathcal{K} \cup \{\text{all}\}$, L inner folds

split the groups into outer folds $\{(T_o, V_o)\}_{o=1}^{J}$
for $o = 1, \ldots, J$ **do**
  split the groups of T_o into inner folds $\{(T_{o\ell}, V_{o\ell})\}_{\ell=1}^{L}$
  for all $\theta = (k, C) \in \Theta$ **and** $\ell = 1, \ldots, L$ **do**
    $S \leftarrow S_k(\mathcal{D}_{T_{o\ell}})$; standardize and fit (3) on $T_{o\ell}$
    $a_{o\ell}(\theta) \leftarrow$ accuracy on $V_{o\ell}$
  end for
  $\theta_o^\star = (k^\star, C^\star) \leftarrow \arg\max_\theta \frac{1}{L} \sum_\ell a_{o\ell}(\theta)$
  $S \leftarrow S_{k^\star}(\mathcal{D}_{T_o})$; standardize and fit (3) on T_o with $C^\star$
  $a_o \leftarrow$ accuracy on V_o ▷ the only use of V_o
end for
return $\hat{A} = \frac{1}{J} \sum_o a_o$, fold SD, and $\{\theta_o^\star\}$

examples and $n_T = |T|$. For voxel j and class c, let $F_j^{(c)}(\mathcal{D}_T)$ be the one-way ANOVA statistic contrasting class c with all other classes on the training examples,

$$F_j^{(c)} = \frac{n_c(\bar{x}_{j,c} - \bar{x}_j)^2 + n_{\bar{c}}(\bar{x}_{j,\bar{c}} - \bar{x}_j)^2}{\frac{1}{n_T - 2} \sum_{i \in T} \left(x_{ij} - \bar{x}_{j,c(i)}\right)^2}, \tag{1}$$

where $\bar{x}_{j,c}$, $\bar{x}_{j,\bar{c}}$ and $\bar{x}_j$ are the means of voxel j over training examples in class c, outside it, and overall, and $c(i)$ is the side of the contrast containing example i. The score and the selected set are

$$s_j(\mathcal{D}_T) = \sum_{c=1}^{M} F_j^{(c)}(\mathcal{D}_T), \qquad S_k(\mathcal{D}_T) = \text{top}_k \left\{ s_j(\mathcal{D}_T) \right\}_{j=1}^{p}, \tag{2}$$

with S_{all} the full candidate pool; for $M = 2$, $s_j = 2t_j^2$. The score uses no spatial information: a voxel's position plays no part in whether it is selected. Selected voxels are standardized with training-fold statistics and one ℓ_2-penalized logistic regression is fitted per class,

$$(\mathbf{w}_c, b_c) = \arg\min_{\mathbf{w}, b} \ \tfrac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_{i \in T} \log\!\left(1 + e^{-\tilde{y}_{ic}(\mathbf{w}^\top \mathbf{z}_i + b)}\right), \tag{3}$$

with $\tilde{y}_{ic} = \pm 1$ according to whether $y_i = c$; prediction is $\hat{y} = \arg\max_c \ \mathbf{w}_c^\top \mathbf{z} + b_c$, and the hyperparameters are $\theta = (k, C)$.

2.3 Evaluation protocol and arms

Every reported accuracy is the nested estimate of Algorithm 1. For Haxby the outer folds are leave-one-run-out (LORO; $J = 12$, 11 for subject 5) with $L = 4$ run-grouped inner folds and $C \in \{0.1, 1, 10\}$; for StarPlus, $J = 10$ trial-grouped folds, $L = 6$, $\mathcal{K} = \{10, \ldots, 1000\}$ and $C \in \{10^{-3}, \ldots, 10^2\}$. Including $k = $ all lets the inner loop decline selection. Each candidate pool is evaluated with and without the selection step on identical folds:

- **Prior (VT)**: the VT mask, $\mathcal{K} = \{50, 100, 200, 300, 500\}$. Without selection this is the standard analysis of these data (Haxby et al., 2001, 2011).

- **No prior (whole brain)**: all 39,912 voxels, $\mathcal{K} = \{100, 500, 1000, 2000, 5000\}$. So that the no-selection arm is not handicapped by too weak a penalty, it is repeated with C extended to 10^{-4}.
- **Prior removed (whole brain minus VT)**: every VT voxel removed, with selection and the whole-brain grid.

VT membership is fixed in advance and independent of the labels, so restricting to it or excluding it is not a source of leakage.

Other controls and classifiers. To test whether the comparison depends on the selection statistic, on the use of the labels or on the classifier, the whole-brain and VT pools are also evaluated with five further arms: the multi-class one-way ANOVA F ranking of `scikit-learn`'s `SelectKBest` in place of Eq. (2); three reductions that do not use the labels – k voxels drawn at random, the k highest-variance voxels, and the first $k \in \{20, 50, 100, 200, 400\}$ principal components of the standardized pool; and the selection of Eq. (2) followed by a linear or a Gaussian-kernel support vector machine (SVM) instead of logistic regression. These arms use $C = 1$ and the grid $\mathcal{K}$ of their pool without the option $k =$ all, so that every control must reduce the pool; rankings, variances and projections are computed on training runs only. The Gaussian-kernel SVM is evaluated twice: with `scikit-learn`'s default kernel width, $\gamma_0 = 1/(d\,\hat{\sigma}^2)$ for d standardized input features of pooled variance $\hat{\sigma}^2$, and with $C \in \{0.1, 1, 10, 100, 1000\}$ and the width $\gamma \in \{0.01, 0.1, 1, 10\} \times \gamma_0$ tuned in the inner loop together with k.

2.4 Where selection puts its voxels

For every subject and outer fold of the whole-brain analysis, voxels are ranked by Eq. (2) on that fold's training runs – exactly the ranking Algorithm 1 uses, so no map below involves held-out data – and the top $k \in \{500, 1000\}$ (the values the inner loop chose in almost every fold) are located relative to the prior. For the selected set S we report its share inside VT, the enrichment

$$E = \frac{|S \cap \mathrm{VT}|/k}{|\mathrm{VT}|/p}, \tag{4}$$

and the one-sided hypergeometric probability of at least $|S \cap \mathrm{VT}|$ VT voxels in a random draw of k from p. Selected voxels outside VT are characterized by their distance to the nearest VT voxel (in scanner coordinates) and by their position along the anterior–posterior axis, normalized to the extent of the whole-brain mask; the fraction of folds in which each voxel is selected gives a per-subject stability map. Because the data are in native space without a common template, we report per-subject maps and statistics rather than a group map.

2.5 Statistical inference

Permutation tests of the procedure. The pipeline is re-run under N label permutations π_b, and $p = \big(1 + \sum_b \mathbb{1}[\hat{A}(\pi_b(\mathbf{y})) \geq \hat{A}(\mathbf{y})]\big)/(N+1)$ (Ojala and Garriga, 2010). For Haxby, π_b permutes the category labels of the stimulus blocks within each run, preserving block structure and run membership – the exchangeability-respecting null for a block design under LORO (Etzel and Braver, 2013; Stelzer et al., 2013). For the VT pool ($N = 50$) the full nested procedure, inner search included, is repeated. For the whole brain minus VT ($N = 50$) the nested search is replaced, for computational reasons, by the hyperparameters it chose most often for that pool ($k = 1{,}000$, $C = 1$), applied identically to real and permuted labels. For StarPlus ($N = 50$) labels are permuted across examples with trial grouping retained.

Comparing arms. Arms evaluated on identical folds give per-fold differences d_o; we use the Nadeau–Bengio corrected resampled t-test, $t = \bar{d}/\sqrt{(1/J + n_V/n_T)\,\hat{\sigma}_d^2}$ with $J-1$ degrees of freedom, because cross-validation training sets overlap. To show that two arms are *equivalent*, subject-level differences are tested with two one-sided tests (TOST) against margins of ± 3 and ± 5 percentage points.

2.6 Leaky protocols

$\hat{A}$ estimates the accuracy of the procedure only if everything learned in outer fold o is a function of T_o alone and the test examples are independent of it. *Selection before the split* ranks voxels once on all data, $S_k(\mathcal{D})$ instead of $S_k(\mathcal{D}_{T_o})$, and then reports the best (k, C) found on the evaluation folds (Ambroise and McLachlan, 2002; Cawley and Talbot, 2010); under the null its bias grows with p/n, since the largest of p null F statistics grows roughly as $2 \log p$. *Ungrouped cross-validation* places temporally adjacent volumes of one block in both training and test sets, where they share the label and, through autocorrelation, the noise. We score these protocols on real data and on block-permuted labels, where the true accuracy is chance: the `nilearn` multi-class example (`plot_haxby_multiclass.py`: `SelectKBest`(F, $k = 500$), linear SVM, stratified `cv=5`), as distributed with `nilearn` releases up to and including 0.14.1 and with LORO folds, and selection before the split on the whole brain and on VT. A ground-truth simulation isolates the two mechanisms: pure Gaussian noise in the Haxby layout (12 runs of 8 blocks of 9 volumes; p from 300 to 40,000), each voxel an independent AR(1) process within runs with $\phi \in \{0, 0.5, 0.8\}$, plus a beta-series layout with one pattern per block, over 5 seeds.

2.7 Software and use of AI assistance

Analyses used Python 3.12.13, `scikit-learn` 1.6.1 (Pedregosa et al., 2011), `nilearn` 0.13.1 (Abraham et al., 2014) and `numpy` 2.0.2 on CPU; the null-label, whole-brain and feature-location analyses and the comparison of controls and classifiers ran under `scikit-learn` 1.6.1 / `nilearn` 0.14.1. Checks across library versions are reported in Appendix E. During the preparation of this work, the author used Claude (Anthropic) for writing and formatting of the paper. The author reviewed and edited the output as needed and takes full responsibility for the content of the article.

3 Results

3.1 Selection helps only when there is no prior

With the prior, selection adds nothing. Within VT (Table 1; Fig. 1A), the pipeline averages 0.714 with selection and 0.718 without it (-0.4 points; per-subject differences -1.6 to $+0.0$ points against fold-level standard deviations of 0.07–0.12), and the inner loop declined selection in 50 of 71 outer folds. Every subject reaches the permutation floor $p = 0.0196$ ($N = 50$; null means 0.122–0.128, chance 0.125). The same holds on StarPlus (Table 2; Fig. 1B): beta-series features reach 0.925 with selection and 0.942 without it (-1.7 points), the no-selection arm itself significant in every subject ($p \leq 0.0196$, $N = 50$). No subject on either dataset shows a significant difference between the arms (all corrected $p \geq 0.24$); pooling the 9 subjects, selection changes accuracy by -0.8 points (90% CI -2.7 to $+1.0$), and TOST establishes equivalence within ± 3 points ($p = 0.031$) and ± 5 points ($p = 0.002$).

Without the prior, selection is essential. With all 39,912 voxels as candidates (Table 3), the pipeline without selection reaches only 0.434, whereas with selection it reaches 0.688: a gain of

Table 1: Haxby, prior (VT) pool: nested LORO accuracy without and with the selection step (mean ± SD over outer folds); the no-selection column is the standard analysis of these data (Haxby et al., 2001, 2011). Permutation test of the full nested procedure with block-wise within-run label permutation ($N = 50$; chance = 0.125).

Subject	VT voxels	No selection (all VT)	+ Selection	Δ (pts)	Null mean	p
1	577	0.791 ± 0.089	0.774 ± 0.085	−1.6	0.127	0.0196
2	464	0.708 ± 0.092	0.708 ± 0.092	+0.0	0.126	0.0196
3	307	0.697 ± 0.117	0.690 ± 0.115	−0.7	0.122	0.0196
4	675	0.557 ± 0.099	0.556 ± 0.095	−0.1	0.126	0.0196
5	422	0.782 ± 0.069	0.782 ± 0.069	+0.0	0.128	0.0196
6	348	0.775 ± 0.076	0.775 ± 0.076	+0.0	0.125	0.0196
Mean		**0.718**	**0.714**	−0.4		

Table 2: StarPlus: beta-series features without and with selection (mean ± SD over 10 outer folds); p: permutation test of the no-selection arm repeating its full nested procedure ($N = 50$; chance = 0.5).

Subject	Voxels	No selection	+ Selection	Δ (pts)	p
04799	4949	0.975 ± 0.053	0.925 ± 0.087	−5.0	0.0196
04820	5015	0.900 ± 0.079	0.850 ± 0.115	−5.0	0.0196
04847	4691	0.950 ± 0.087	1.000 ± 0.000	+5.0	0.0196
Mean		**0.942**	**0.925**	−1.7	

+25.4 points, positive in 6 of 6 subjects and significant in each (corrected $p \leq 0.034$). The inner loop chose a strict subset in all 71 outer folds. The gap is not an artifact of an under-regularized baseline: allowed C down to 10^{-4}, the no-selection arm never chose a value below 0.1 and its accuracies were unchanged (maximum difference 0.0 points). Whole-brain selection recovers most of what the prior provides (0.688 versus 0.718), and in subjects 3, 4 and 6 it exceeds it. The two are interchangeable controls of dimensionality: once either has reduced the pool, the other adds nothing.

Which voxels, not how many. Table 4 repeats both pools with other controls and classifiers. The whole-brain gain is not specific to the ranking statistic or to the classifier: the multi-class ANOVA F gives +23.7 points against +24.1 for the summed score (+25.4 in the main analysis, which also tunes C), a linear SVM gains +24.6 and a Gaussian-kernel SVM +38.0, each in 6 of 6 subjects. It does depend on using the labels. A random subset of voxels lowers whole-brain accuracy by 7.1 points and the highest-variance voxels lower it by 18.1 (both in every subject), and principal components add only 2.7 (4 of 6 subjects). Reducing the number of voxels is therefore not what helps; selection, like the prior, helps because it locates the informative ones. Within the prior the choice of control hardly matters for the linear classifiers: required to discard voxels, every reduction lowers the mean accuracy, by 1.0 to 2.1 points. For the two label-based rankings, the random subset and the linear SVM this is equivalent to no change within ±3 points (TOST $p \leq 0.020$); the variance filter and principal components vary more between subjects and are equivalent within ±5 points only ($p \leq 0.041$; at ±3, $p = 0.13$ and $p = 0.27$).

A nonlinear decoder. The Gaussian-kernel SVM shows that the equivalence within the prior depends on the decoder being tuned. With default hyperparameters it is a much weaker decoder of these data (0.492 within VT, 22.2 points below logistic regression), and selection within VT raises

it by +4.0 points on average – by 5.6–6.3 points in 4 subjects and not at all in the other 2 – which is not equivalent to zero at either margin (TOST $p = 0.77$ at ± 3). With C and the kernel width tuned in the inner loop it reaches 0.662 without selection, and selection then changes its accuracy by +0.4 points (per subject −1.4 to +1.8; 90% CI −0.6 to +1.3), equivalent to zero within ±3 points (TOST $p = 0.001$). On the whole brain the tuned SVM still depends on selection (0.277 without, 0.621 with). The gain of the untuned SVM within the prior is therefore best read as compensation for poor hyperparameters rather than information that selection adds. Two caveats apply: the tuned SVM remains below logistic regression within VT (−5.2 points), and the inner loop chose the widest kernel of the grid in 128 of 142 cases, so its optimum may lie at or beyond the edge of the grid.

Table 3: Haxby, no-prior pool (39,912 voxels): nested LORO accuracy without and with selection (mean ± SD over outer folds).

Subject	No selection	+ Selection	Δ (pts)
1	0.424 ± 0.078	0.763 ± 0.079	+33.9
2	0.448 ± 0.135	0.611 ± 0.127	+16.3
3	0.449 ± 0.153	0.748 ± 0.057	+29.9
4	0.397 ± 0.074	0.625 ± 0.098	+22.8
5	0.352 ± 0.150	0.596 ± 0.129	+24.4
6	0.534 ± 0.128	0.784 ± 0.118	+25.0
Mean	**0.434**	**0.688**	+25.4

Table 4: Haxby: other dimensionality controls and classifiers under the nested LORO protocol (mean accuracy over six subjects; $C = 1$ except in the tuned rows; k chosen in the inner loop from a grid without the option "all"). Δ: difference in points from no reduction with the same classifier and pool; in parentheses, the number of subjects in which it is positive. Per-subject values are in Table 11.

		Whole brain (39,912 voxels)		Prior (VT)	
Classifier	Dimensionality control	Accuracy	Δ (pts)	Accuracy	Δ (pts)
Logistic regression	None (all voxels)	0.435	–	0.714	–
	Summed one-vs-rest F (Eq. 2)	0.676	+24.1 (6/6)	0.702	−1.2 (0/6)
	Multi-class ANOVA F	0.672	+23.7 (6/6)	0.702	−1.2 (0/6)
	Random voxels	0.364	−7.1 (0/6)	0.698	−1.6 (0/6)
	Highest-variance voxels	0.253	−18.1 (0/6)	0.698	−1.6 (2/6)
	Principal components	0.462	+2.7 (4/6)	0.693	−2.1 (1/6)
Linear SVM	None (all voxels)	0.425	–	0.689	–
	Summed one-vs-rest F (Eq. 2)	0.671	+24.6 (6/6)	0.679	−1.0 (1/6)
Gaussian-kernel SVM	None (all voxels)	0.169	–	0.492	–
	Summed one-vs-rest F (Eq. 2)	0.549	+38.0 (6/6)	0.532	+4.0 (4/6)
Gaussian-kernel SVM, tuned	None (all voxels)	0.277	–	0.662	–
	Summed one-vs-rest F (Eq. 2)	0.621	+34.4 (6/6)	0.666	+0.4 (4/6)

3.2 Selection rediscovers the prior

Table 5 and Fig. 2A,B show where the whole-brain selection places its voxels. Although VT makes up only 0.8–1.7% of the candidate pool, it contained 15–33% of the top 500 voxels – an enrichment of 9–

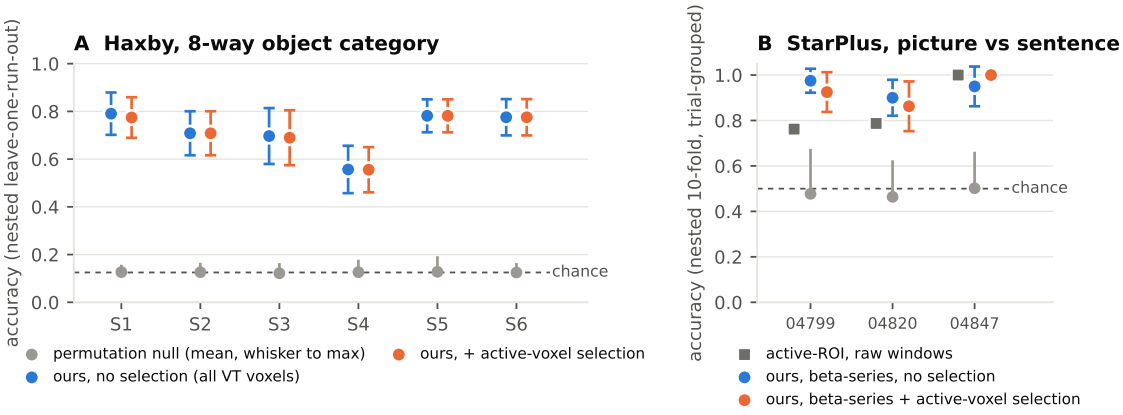

Figure 1: Per-subject nested cross-validation accuracy (mean ± SD over outer folds). (A) Haxby, prior (VT) pool: all VT voxels (blue) and selection within VT (orange); gray, permutation null of the full nested procedure (mean, whisker to maximum over 50 permutations); dashed line, chance (1/8). (B) StarPlus, beta-series features without (blue) and with (orange) selection, with the active-ROI method of Chowdhury (2014) on raw windows (gray squares); light gray, permutation null (50 permutations); dashed line, chance (1/2).

35-fold (mean 21) – with no fold of any subject below 12% and every fold significant (hypergeometric $p < 10^{-41}$). The convergence holds at $k = 1000$ (11–24% in VT; 8–22-fold) and is stable: 152–340 voxels per subject were selected in every outer fold, out of 778–1192 selected in any. Of the selected voxels, 0.6–4.3% fell in the face-selective mask and 4.5–20.7% in the house-selective mask distributed with the data. This explains the interchangeability of Sect. 3.1: the label statistic and the anatomical prior single out largely the same voxels, so applying one after the other adds little.

The overlap is not complete. Most selected voxels lie outside VT (67–85%), and they are not a rim of spillover around it: only 15–21% lie within 10 mm of a VT voxel (median distance 19–39 mm), and 42–80% lie in the posterior third of the brain. The maps (Fig. 2A) show one posterior occipitotemporal territory in every subject – VT together with occipital cortex posterior and dorsal to it – most compact in the subjects with the highest accuracy and more scattered in the weakest.

3.3 The prior is sufficient but not necessary

With VT removed from the candidates, nested accuracy remained far above chance in every subject (0.475–0.740, mean 0.606; Table 5, Fig. 2C). The fixed-hyperparameter version of the same pipeline averaged 0.593, against permutation null means of 0.120–0.129 (largest null accuracy 0.220), so every subject reached $p \leq 0.0196$. Without VT, accuracy differed by −11.2 points from the prior-only analysis and by −8.2 points from whole-brain selection with VT available. Per voxel, the prior is the most informative region – its few hundred voxels matched or outperformed the best thousand or so of the remaining 39,000 in 5 of 6 subjects – but it is not the only one. The cost of removing it split the subjects cleanly: at most 3.6 points in subjects 3, 4 and 6, the same subjects in which whole-brain selection exceeded the prior, against 17–31 points in the others. Where an anatomical prior is drawn too narrowly for an individual, data-driven selection can recover the signal it leaves out.

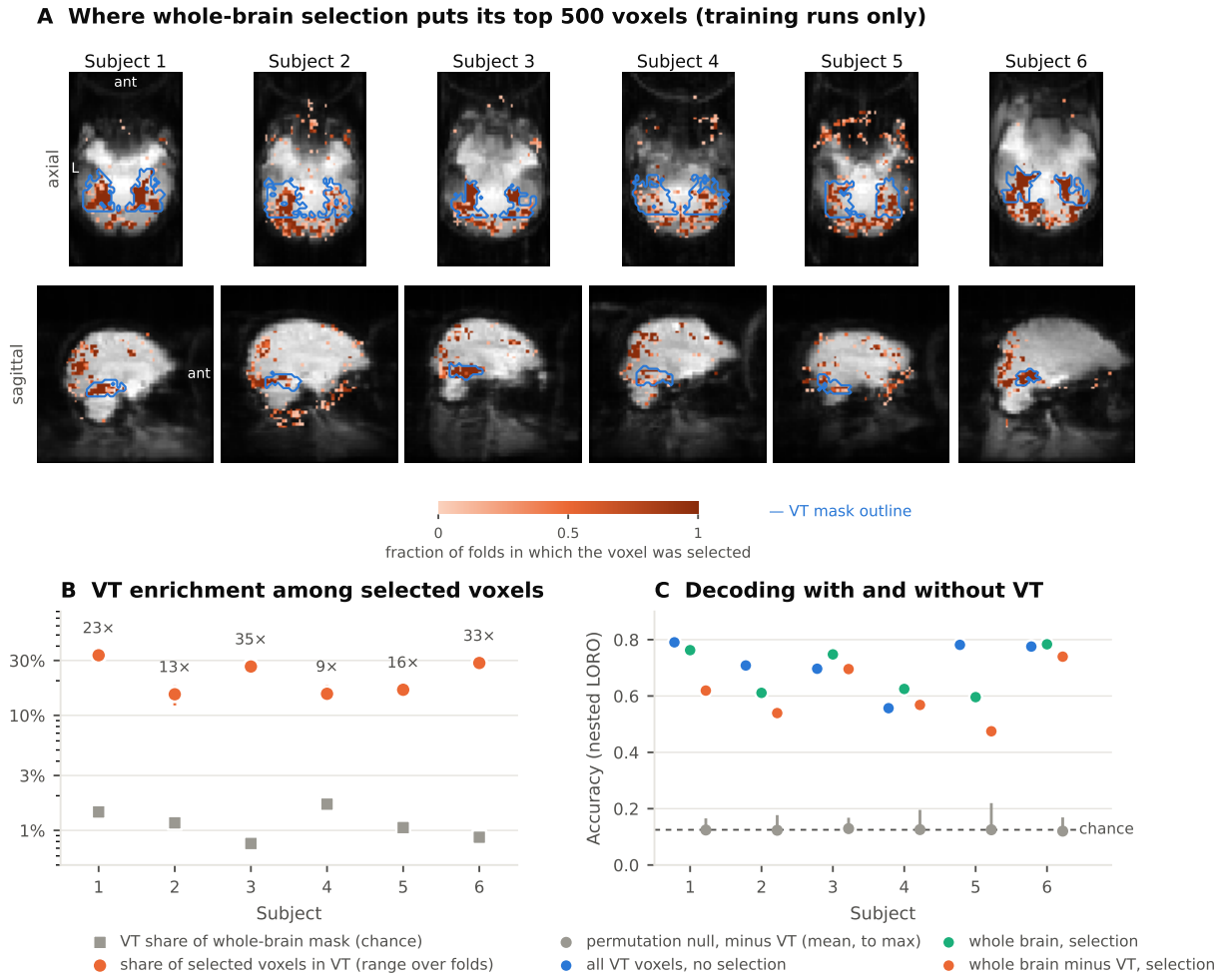

Figure 2: Where data-driven selection finds the signal (Haxby, six subjects). (A) Fraction of the leave-one-run-out folds in which each voxel was among the 500 highest-scoring of 39,912 (Eq. 2, training runs only), on axial (top; anterior up) and sagittal (bottom; anterior right) slices of each subject's mean functional image; blue outline, the VT prior. (B) Share of the selected voxels inside VT (orange; line, range over folds) against VT's share of the candidate pool (gray squares); labels give the enrichment. (C) Nested LORO accuracy from all VT voxels without selection (blue), from whole-brain selection (green) and from whole-brain selection with VT removed (orange); gray, permutation null of the minus-VT pipeline with fixed hyperparameters (mean, whisker to maximum); dashed line, chance (1/8).

Table 5: Where whole-brain selection puts its voxels (top 500 voxels, ranked on the training runs of each LORO fold; means over folds) and decoding with the prior removed. VT share: VT voxels as a percentage of the 39,912 candidates (the chance level for the next column). Posterior: percentage of the selected voxels outside VT that lie in the posterior third of the brain. Minus VT: nested LORO accuracy with VT removed from the candidates (mean ± SD over folds); null mean and permutation p are for the same pipeline with its modal hyperparameters (k = 1,000, $C = 1$) fixed ($N = 50$; chance = 0.125).

		Selected voxels				Decoding, minus VT		
Subj.	VT voxels	VT share (%)	In VT (%)	Enrichment	Posterior (%)	Accuracy	Null	p
1	577	1.45	33.3	23×	80	0.619 ± 0.101	0.124	0.020
2	464	1.16	15.2	13×	57	0.539 ± 0.166	0.123	0.020
3	307	0.77	26.6	35×	75	0.696 ± 0.051	0.129	0.020
4	675	1.69	15.4	9×	74	0.568 ± 0.085	0.126	0.020
5	422	1.06	16.7	16×	42	0.475 ± 0.125	0.125	0.020
6	348	0.87	28.5	33×	78	0.740 ± 0.064	0.120	0.020

3.4 What leaky protocols report when there is nothing to find

Table 6 scores each protocol on Haxby data with block-permuted labels ($N = 20$ per subject), where the true accuracy is 0.125. The run-grouped `nilearn` example is unbiased (0.127), whereas the example as released, with ungrouped five-fold cross-validation, reports 0.212 – 8.7 points above chance on data with no stimulus information, learned from the temporal similarity of volumes within a block. Selection before the split reports 0.223 on the whole brain and 0.155 within VT. On StarPlus, where 80 patterns face ∼5,000 voxels, it reports 0.925–0.952 on random labels (chance 0.5, $N = 50$) – within a few points of the 100% it reported on the real labels. On real data the same protocol overstated nested accuracy by up to 44.8 points, enough to reverse a comparison with the baseline (Appendix A).

The simulation (Fig. 3) separates the two mechanisms. The correct protocol stays within sampling noise of chance in every condition (0.11–0.18). Selection before the split scales with the number of candidate voxels: on noise in the raw-frame layout it reports 0.396 with 2,000 candidates and 0.728 with 40,000 (0.930 at $\phi = 0.5$), and in the beta-series layout it reaches 1.000 from 10,000 candidates onward. Ungrouped cross-validation is insensitive to dimensionality and driven by autocorrelation: harmless for independent volumes (0.129) but 0.254 at $\phi = 0.8$. The simulated selection bias on raw frames exceeds that observed on real data, as expected: spatially smoothed voxels are strongly correlated, so their effective dimensionality is far below the nominal count. A map of "informative" voxels produced by selection before the split inherits the same bias and can show them anywhere, which is why every selection in Sect. 3.2 was made on training runs only.

How large is the selection bias? For independent features it can be predicted in closed form (Appendix D). Take two balanced classes $y_i = \pm 1$, n examples and p noise features, and let $T_j = n^{-1/2} \sum_i y_i x_{ij}$ be the statistic by which features are ranked. Conditional on T_j, $x_{ij} = y_i T_j/\sqrt{n} + e_{ij}$ with $\sum_i y_i e_{ij} = 0$: a noise feature chosen for its large $|T_j|$ behaves like a feature with a true class difference of $2T_j/\sqrt{n}$. Selecting k of p features on all the data therefore manufactures a signal, and a nearest-centroid classifier cross-validated on those features with n_T training and n_V test examples per fold reports

$$A_{\text{leak}} \approx \Phi\left(\frac{\sqrt{k}\,\left(m_k(p)/n - 1/n_T\right)}{\sqrt{m_k(p)/n + n_V/n_T^2}}\right), \tag{5}$$

where Φ is the standard normal distribution function and $m_k(p)$ is the mean of the k largest of p independent χ^2_1 variables, which grows as $2\log(p/k)$. Across 32 simulated configurations (n from 40 to 320, p from 500 to 40,000, k from 10 to 500) Eq. (5) matches the simulated accuracy of the leaky protocol within 1.7 points for the nearest-centroid classifier and within 4.0 points for logistic regression, while selection inside the split stays at chance (Fig. 4). The bias grows with k and with $\log p$ but falls only slowly with n: with 5,000 candidates and $k = 100$ it is still 0.901 at $n = 320$. For the dimensions of StarPlus ($n = 80$, $p \approx 5{,}000$) and $k = 100$ the equation gives 0.994, above the 0.925–0.952 observed on random labels there, as expected when correlated voxels lower the effective p.

Table 6: Haxby: accuracy on the real labels / mean accuracy on block-permuted labels (true accuracy 0.125) for the `nilearn` example as released (ungrouped `cv=5`; releases up to and including 0.14.1) and run-grouped (LORO), and for selection before the split on the whole-brain and VT pools.

	`nilearn` example		Selection before split	
Subject	`cv=5`	LORO	Whole brain	VT
1	0.590 / 0.222	0.547 / 0.126	0.819 / 0.221	0.793 / 0.155
2	0.535 / 0.221	0.372 / 0.129	0.752 / 0.228	0.716 / 0.160
3	0.604 / 0.214	0.553 / 0.127	0.789 / 0.232	0.697 / 0.154
4	0.499 / 0.239	0.456 / 0.126	0.690 / 0.234	0.581 / 0.163
5	0.488 / 0.205	0.431 / 0.125	0.686 / 0.214	0.783 / 0.150
6	0.633 / 0.172	0.650 / 0.127	0.843 / 0.207	0.787 / 0.146
Null mean	**0.212**	0.127	**0.223**	0.155

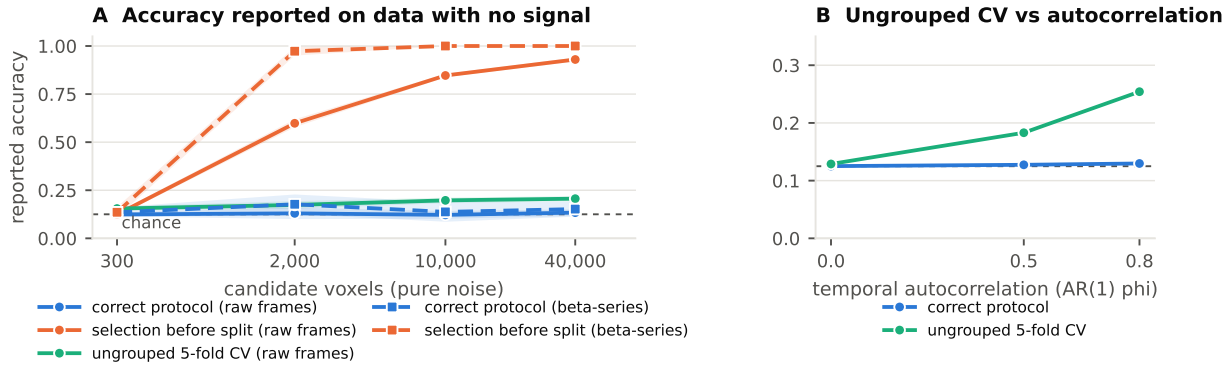

Figure 3: Ground-truth simulation on pure noise in the Haxby layout (12 runs × 8 blocks), mean ± SD over 5 seeds; chance = 1/8. (A) Accuracy reported by each protocol as the number of candidate voxels grows, for raw frames (864 volumes, AR(1) $\phi = 0.5$; circles) and beta-series (96 patterns; squares). (B) Ungrouped five-fold cross-validation versus the correct protocol as temporal autocorrelation increases (averaged over voxel counts).

4 Discussion

Substitutes, not complements. A label-based ranking that knows nothing about anatomy, refit on every training set, chose voxels from the anatomical prior 9–35 times more often than chance on average, and significantly more often in every fold of every subject. The prior and the selection step therefore pick largely the same voxels, which is why applying one after the other adds nothing (equivalence within ±3 points), while applying neither leaves the classifier with 40,000 mostly uninformative voxels (0.434 against 0.688). For practice this suggests a working rule, to be confirmed on other datasets and priors: a selection step is worth its cost when no defensible prior is available, and adds little when one is. Strong ℓ_2 regularization is not a substitute for either: on 40,000 voxels no value of C down to 10^{-4} closed the gap. Nor is a reduction that ignores the labels: random subsets, a variance filter and principal components reduce the whole-brain pool to at most 5,000 dimensions and recover almost none of the gain (Table 4). What the two effective controls share is information about where the signal is, one from the labels and one from prior knowledge of the anatomy. The result is consistent with Chu et al. (2012), who found no benefit of univariate selection for whole-brain support vector machines on structural data, and qualifies it: on functional data without a prior, selection is essential.

Information beyond VT, and priors that are too narrow. The prior is sufficient but not necessary. The voxels outside it decoded at 0.606 (chance 0.125); the selected ones lay away from VT's border rather than in a rim around it, many of them in occipital cortex posterior and dorsal to VT, and in three of six subjects data-driven selection beat the prior while removing the prior cost almost nothing. This agrees with the view that category information in occipitotemporal cortex is partially distributed and extends beyond the classical category-selective regions (Haxby et al., 2001; Spiridon and Kanwisher, 2002; O'Toole et al., 2005), and with object-selective responses in lateral occipital cortex (Malach et al., 1995). A fixed mask encodes population-level knowledge that may fit some individuals poorly, just as the voxel count that decodes best varies between individuals (Bleichner et al., 2014); selection nested inside the evaluation adapts to the individual at no risk of bias. We do not claim that the voxels outside VT represent object category as such: the stimuli differ in low-level image properties between categories, which early visual cortex can exploit (Rice et al., 2014). The claim is about where category-*discriminative* signal lies for a linear decoder.

Evaluation. Both leaky protocols produce large, confident numbers on null data, and they fail for different reasons – one through dimensionality, one through autocorrelation – so fixing one does not fix the other. Three checks follow from the measurements and cost little: (i) state where every label-dependent step, selection included, sits relative to the split, and whether folds respect the temporal grouping of the data; (ii) run the complete pipeline once on permuted labels and report the accuracy it gives, which exposes both errors at once; (iii) when a result was obtained with selection outside the split, evaluate Eq. (5) for its n, p and k – if the equation alone predicts an accuracy near the reported one, the result carries no evidence of signal. Most of the difference between the published tutorial and the standard analysis on the same data is the anatomical prior itself (Appendix B): ungrouped folds inflate the tutorial's estimate, and the whole-brain candidate pool depresses it. After the preprint of this work was cited in the `nilearn` issue tracker, the example was changed to run-grouped cross-validation (Nilearn contributors, 2026a,b); the change is in the development version, and releases up to and including 0.14.1 distribute the ungrouped version measured here.

Limitations. Both selection statistics are univariate filters; multivariate selection, such as recursive elimination (De Martino et al., 2008), could exploit voxels differently. The equivalence within the prior holds for the linear classifiers and for a tuned Gaussian-kernel SVM, but not for the same SVM at default hyperparameters, and the tuned SVM chose the widest kernel available in most

folds; decoders more flexible than these, such as deep networks, were not tested. The comparison of controls and classifiers (Table 4) covers Haxby only. The evidence comes from two datasets and nine subjects; nested estimates over 10–12 folds have standard deviations of about 0.1, so per-subject differences of a few points are not individually reliable (Varoquaux, 2018), and our claims rest on effects consistent across subjects, on equivalence tests and on permutation tests whose floor ($1/(N+1)$) is set by compute. The prior-removed permutation test used fixed rather than nested hyperparameters. The feature-location analysis is per subject in native space with masks defined by the original investigators; no group-level map is reported, and position is summarized along coarse axes rather than by cytoarchitectonic or retinotopic area, which would require anatomical images (absent for one subject) and a functional localizer. The StarPlus results of the original active-ROI method were reproduced for one of three subjects only (Appendix C).

5 Conclusion

Under leakage-free, group-respecting nested evaluation, voxel selection and an anatomical prior are interchangeable controls of dimensionality for fMRI decoding with linear and tuned kernel classifiers: selection is essential without the prior (+25.4 points), equivalent to none with it, and, without seeing any spatial information, concentrates on the very region the prior encodes. Reductions that ignore the labels do not substitute for either. The prior is sufficient but not necessary, and in some individuals data-driven selection outperforms it. Maps of where a decoder finds its signal, like the accuracies they accompany, are meaningful only when selection is nested inside the evaluation; the two common shortcuts measured here report 21% on an eight-class and 95% on a two-class problem with no signal at all.

Acknowledgments

The author thanks the original investigators of the Haxby et al. and StarPlus datasets for making their data publicly available.

Statements and Declarations

Funding. The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests. The author has no relevant financial or non-financial interests to disclose.

Ethics approval. This study is a secondary analysis of publicly available, de-identified human fMRI data from previously published studies (Haxby et al., 2001; Mitchell et al., 2004), collected with the original investigators' ethical approval. No new data were collected from human or animal subjects, and no additional ethical approval was required.

Consent to participate. Not applicable to this secondary analysis; informed consent was obtained by the original studies.

Consent to publish. Not applicable.

Author contributions. T.C.: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, writing – original draft, writing – review and editing.

Data availability. Both datasets are public. The Haxby et al. data (Haxby et al., 2001) are distributed by the PyMVPA project (Hanke et al., 2009) (`http://data.pymvpa.org/datasets/`

haxby2001/) and were obtained through `nilearn.datasets.fetch_haxby` (Abraham et al., 2014); the StarPlus data (Mitchell et al., 2004) are distributed by the Center for Cognitive Brain Imaging, Carnegie Mellon University (`http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/`).

Code availability. All code needed to reproduce every number, table and figure, including the feature-location analysis, the leaky protocols, the simulation and the closed-form check, is available at `https://github.com/tanayz/active-voxel-fmri-decoding` (MIT license). A preprint of an earlier version of this work is available at Research Square (`https://doi.org/10.21203/rs.3.rs-10828698/v1`).

References

Hunar Abdulrahman and Richard N. Henson. Effect of trial-to-trial variability on optimal event-related fMRI design: Implications for beta-series correlation and multi-voxel pattern analysis. *NeuroImage*, 125:756–766, 2016. doi: 10.1016/j.neuroimage.2015.11.009.

Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8:14, 2014. doi: 10.3389/fninf.2014.00014.

Christophe Ambroise and Geoffrey J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99(10):6562–6566, 2002. doi: 10.1073/pnas.102102699.

Martin G. Bleichner, Johan M. Jansma, Jim Sellmeijer, Mathijs Raemaekers, and Nicolas F. Ramsey. Give me a sign: decoding complex coordinated hand movements using high-field fMRI. *Brain Topography*, 27(2):248–257, 2014. doi: 10.1007/s10548-013-0322-x.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

Tanay Chowdhury. Enhanced predictor of human cognitive state. M.S. thesis, University of Illinois at Chicago, 2014.

Carlton Chu, Ai-Ling Hsu, Kun-Hsien Chou, Peter Bandettini, and ChingPo Lin. Does feature selection improve classification accuracy? impact of sample size and feature selection on classification using anatomical magnetic resonance images. *NeuroImage*, 60(1):59–70, 2012. doi: 10.1016/j.neuroimage.2011.11.066.

Federico De Martino, Giancarlo Valente, Noël Staeren, John Ashburner, Rainer Goebel, and Elia Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, 43(1):44–58, 2008. doi: 10.1016/j.neuroimage.2008.06.037.

Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998. doi: 10.1038/33402.

Joset A. Etzel and Todd S. Braver. MVPA permutation schemes: Permutation testing in the land of cross-validation. In *2013 International Workshop on Pattern Recognition in Neuroimaging*, pages 140–143, 2013. doi: 10.1109/prni.2013.44.

Kalanit Grill-Spector and Kevin S. Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548, 2014. doi: 10.1038/nrn3747.

J. Swaroop Guntupalli, Michael Hanke, Yaroslav O. Halchenko, Andrew C. Connolly, Peter J. Ramadge, and James V. Haxby. A model of representational spaces in human cortex. *Cerebral Cortex*, 26(6):2919–2934, 2016. doi: 10.1093/cercor/bhw068.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

Michael Hanke, Yaroslav O. Halchenko, Per B. Sederberg, Stephen José Hanson, James V. Haxby, and Stefan Pollmann. PyMVPA: a Python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, 7(1):37–53, 2009. doi: 10.1007/s12021-008-9041-y.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009.

James V. Haxby, M. Ida Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001. doi: 10.1126/science.1063736.

James V. Haxby, J. Swaroop Guntupalli, Andrew C. Connolly, Yaroslav O. Halchenko, Bryan R. Conroy, M. Ida Gobbini, Michael Hanke, and Peter J. Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

James V. Haxby, Andrew C. Connolly, and J. Swaroop Guntupalli. Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience*, 37(1):435–456, 2014. doi: 10.1146/annurev-neuro-062012-170325.

John-Dylan Haynes and Geraint Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534, 2006. doi: 10.1038/nrn1931.

Martin N. Hebart and Chris I. Baker. Deconstructing multivariate decoding for the study of brain function. *NeuroImage*, 180:4–18, 2018. doi: 10.1016/j.neuroimage.2017.08.005.

Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17(11): 4302–4311, 1997. doi: 10.1523/jneurosci.17-11-04302.1997.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining. *ACM Transactions on Knowledge Discovery from Data*, 6(4):1–21, 2012. doi: 10.1145/2382577.2382579.

Nikolaus Kriegeskorte, Marieke Mur, Douglas A. Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A. Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008. doi: 10.1016/j.neuron.2008.10.043.

Nikolaus Kriegeskorte, W Kyle Simmons, Patrick S F Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12(5): 535–540, 2009. doi: 10.1038/nn.2303.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/s0047-259x(03) 00096-4.

R. Malach, J. B. Reppas, R. R. Benson, K. K. Kwong, H. Jiang, W. A. Kennedy, P. J. Ledden, T. J. Brady, B. R. Rosen, and R. B. Tootell. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, 92(18):8135–8139, 1995. doi: 10.1073/pnas.92.18.8135.

Tom M. Mitchell, Rebecca Hutchinson, Radu S. Niculescu, Francisco Pereira, Xuerui Wang, Marcel Just, and Sharlene Newman. Learning to decode cognitive states from brain images. *Machine Learning*, 57(1-2):145–175, 2004.

Jeanette A. Mumford, Benjamin O. Turner, F. Gregory Ashby, and Russell A. Poldrack. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, 59(3):2636–2643, 2012. doi: 10.1016/j.neuroimage.2011.08.076.

Jeanette A. Mumford, Tyler Davis, and Russell A. Poldrack. The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *NeuroImage*, 103:130–138, 2014. doi: 10.1016/j.neuroimage.2014.09.026.

Nilearn contributors. Leakage in plot_haxby_multiclass feature selection. GitHub issue #6573, `https://github.com/nilearn/nilearn/issues/6573`, 2026a. Opened 8 September 2026.

Nilearn contributors. Use run-aware cross-validation in the Haxby multiclass example. GitHub pull request #6591, `https://github.com/nilearn/nilearn/pull/6591`, 2026b. Merged 24 September 2026.

Kenneth A. Norman, Sean M. Polyn, Greg J. Detre, and James V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–430, 2006. doi: 10.1016/j.tics.2006.07.005.

Markus Ojala and Gemma C. Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11:1833–1863, 2010.

Alice J. O'Toole, Fang Jiang, Hervé Abdi, and James V. Haxby. Partially distributed representations of objects and faces in ventral temporal cortex. *Journal of Cognitive Neuroscience*, 17(4):580–590, 2005. doi: 10.1162/0898929053467550.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, 45(1 Suppl):S199–S209, 2009.

Russell A. Poldrack, Grace Huckins, and Gael Varoquaux. Establishment of best practices for evidence for prediction. *JAMA Psychiatry*, 77(5):534, 2020. doi: 10.1001/jamapsychiatry.2019.3671.

G. E. Rice, D. M. Watson, T. Hartley, and T. J. Andrews. Low-level image properties of visual objects predict patterns of neural response across category-selective regions of the ventral visual pathway. *Journal of Neuroscience*, 34(26):8837–8844, 2014. doi: 10.1523/jneurosci.5265-13.2014.

Jesse Rissman, Adam Gazzaley, and Mark D'Esposito. Measuring functional connectivity during distinct stages of a cognitive task. *NeuroImage*, 23(2):752–763, 2004. doi: 10.1016/j.neuroimage.2004.06.035.

Mona Spiridon and Nancy Kanwisher. How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron*, 35(6):1157–1165, 2002. doi: 10.1016/s0896-6273(02)00877-2.

Johannes Stelzer, Yi Chen, and Robert Turner. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage*, 65:69–82, 2013. doi: 10.1016/j.neuroimage.2012.09.063.

Gaël Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180:68–77, 2018. doi: 10.1016/j.neuroimage.2017.06.061.

Gaël Varoquaux, Pradeep Reddy Raamana, Denis A. Engemann, Andrés Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179, 2017. doi: 10.1016/j.neuroimage.2016.10.038.

A Leakage on real data

Table 7 compares, for the same subject and features, the accuracy of the leaky protocol (voxel ranking and k/C sweep on the full dataset) with the nested result. The inflation scales with the ratio of candidate voxels to examples, as the selection-bias argument predicts: it averages 20.8 points on the Haxby beta-series track (39,912 candidate voxels for 96 patterns), 7 points on StarPlus ($\sim$5,000 voxels for 80 patterns) and 10 points on the Haxby raw-frame track (864 volumes). For Haxby subject 2 the leaky beta-series number (0.823) was above every baseline, whereas the nested number (0.375) is below the run-grouped `nilearn` baseline: the leak did not merely inflate a real effect but reversed a conclusion.

Table 7: Inflation from ranking voxels on the full dataset before cross-validation (leaky) versus nested cross-validation, same features and classifiers. Haxby beta rows: beta-series, whole brain (96 patterns, 39,912 candidate voxels); last row: raw frames, whole brain.

Dataset	Subject	Leaky	Nested	Δ (pts)
StarPlus	04799	1.000	0.925	−7.5
StarPlus	04820	1.000	0.863	−13.7
StarPlus	04847	1.000	1.000	0.0
Haxby (beta)	1	0.760	0.510	−25.0
Haxby (beta)	2	0.823	0.375	−44.8
Haxby (beta)	3	0.875	0.750	−12.5
Haxby (beta)	4	0.823	0.646	−17.7
Haxby (beta)	5	0.966	0.830	−13.6
Haxby (beta)	6	0.812	0.698	−11.4
Haxby (raw)	2	0.720	0.619	−10.1

B From the tutorial to the standard analysis

Table 8 changes one element at a time, from the `nilearn` tutorial pipeline to the standard whole-VT analysis. On subject 2 the example reports 0.535 with ungrouped folds but 0.372 with LORO folds; restricting that same pipeline to the VT mask raises it to 0.683, and across subjects the VT mask alone accounts for nearly the whole difference between the tutorial and the standard analysis, while standardization, the classifier and tuning of C contribute only fold-level noise. The one-versus-one variant of the example drops from 0.472 to 0.344 under LORO. Releases of `nilearn` up to and including 0.14.1 distribute the example with these ungrouped folds; its development version has used six-fold run-grouped cross-validation since 24 September 2026 (Nilearn contributors, 2026b).

Table 8: Haxby, from the `nilearn` tutorial pipeline (`SelectKBest`, $k = 500$, linear SVM) to the standard whole-VT analysis, one change at a time (LORO accuracy). A: the tutorial pipeline with its candidate pool restricted to the VT mask; B: A + feature standardization; C: B with logistic regression at $C = 1$; D: C with C tuned in a nested inner loop (all VT voxels, no selection); E: D + selection within VT.

Subject	A	B	C	D	E
1	0.763	0.782	0.792	0.791	0.774
2	0.683	0.718	0.716	0.708	0.708
3	0.670	0.632	0.672	0.697	0.690
4	0.560	0.551	0.558	0.557	0.556
5	0.753	0.746	0.783	0.782	0.782
6	0.734	0.712	0.763	0.775	0.775
Mean	0.694	0.690	0.714	0.718	0.714

Table 9: StarPlus: beta-series features with selection (nested CV, mean ± SD over 10 outer folds), the active-ROI method of Chowdhury (2014) on raw windows, and a permutation test repeating the full nested procedure ($N = 50$; chance = 0.5).

Subject	Selection	Active-ROI	Null mean	p
04799	0.925 ± 0.087	0.762	0.477	0.0196
04820	0.863 ± 0.109	0.787	0.464	0.0196
04847	1.000 ± 0.000	1.000	0.502	0.0196
Mean	**0.929**	0.850	–	–

C StarPlus: comparison with the active-ROI method

The active-ROI method ranks whole anatomical regions by standalone accuracy after averaging each region's voxels into one feature and adds regions greedily. A reimplementation following its description matches the reported accuracy for subject 04847 (1.00 versus 1.00) but falls 15–20 points short for 04799 and 04820 (0.76 and 0.79 versus 0.92 and 0.96) across every preprocessing variant tried; the exact normalization choices of the original are not fully documented. Single-trial beta-series suit StarPlus's slow, sparse events, whereas on Haxby's contiguous blocks raw frames were better (Table 10), as single-trial estimation theory predicts (Mumford et al., 2012, 2014; Abdulrahman and Henson, 2016).

D The accuracy of selection before the split on noise

Setting. Let $y_i = \pm 1$ be balanced labels of n examples and x_{ij}, $j = 1, \ldots, p$, independent standard normal features, independent of the labels. The leaky protocol ranks features by T_j^2, $T_j = n^{-1/2} \sum_i y_i x_{ij} \sim \mathcal{N}(0, 1)$, computed on all n examples, keeps the set S of the k largest, and cross-validates a classifier on S with folds of n_T training and n_V test examples ($n_T + n_V = n$).

Selected noise is signal. For each feature, $(x_{1j}, \ldots, x_{nj})$ decomposes orthogonally as $x_{ij} = y_i \mu_j + e_{ij}$ with $\mu_j = T_j/\sqrt{n}$ and $\sum_i y_i e_{ij} = 0$, where the residual $e_{\cdot j}$ is independent of T_j. Conditioning on the selection, which depends on the data only through the T_j, leaves the residuals untouched: the selected features have class means $\pm\mu_j$ and nearly unit-variance noise, as if each carried a true effect μ_j. The squared distance between the manufactured class means is $\sum_{j \in S} (2\mu_j)^2 = 4k\, m/n$, with $m = \frac{1}{k} \sum_{j \in S} T_j^2$ the mean of the k largest of p independent χ_1^2 variables. For $k \ll p$, with $q = k/(2p)$ and $c = \Phi^{-1}(1 - q)$, $m \approx 1 + c\, \varphi(c)/q$, where φ is the standard normal density; this grows as $2 \log(p/k)$.

Cross-validated nearest centroid. The classifier estimates the class-mean difference on the training fold, $\hat{d}_j = 2\mu_j + \frac{2}{n_T} \sum_{i \in T} y_i e_{ij}$. Because the residuals sum to zero over all examples, $\sum_{i \in T} y_i e_{ij} = -\sum_{i \in V} y_i e_{ij}$: the training-fold noise is the negative of the test-fold noise, the familiar reason cross-validated accuracy falls slightly below chance on unselected noise. A test example l with $y_l = +1$ has score $\sum_{j \in S} \hat{d}_j (\mu_j + e_{lj})$, with mean $2k\, m/n - 2k/n_T$ and variance approximately $\sum_{j \in S} \hat{d}_j^2 \approx 4k\, m/n + 4k\, n_V/n_T^2$, neglecting the error of the estimated midpoint. A normal approximation to the score gives Eq. (5). Setting $k = p$ and $m = 1$ recovers the below-chance accuracy of cross-validation without selection.

Validation. Fig. 4 compares Eq. (5) with simulation ($J = 10$ stratified folds, 20 seeds per configuration). The mean absolute error over the 32 configurations is 0.5 points for the nearest-centroid classifier (maximum 1.7); standardized ℓ_2 logistic regression with $C = 1$ follows the same

curve within 4.0 points. Selecting inside each training fold gives 0.455–0.544. The equation assumes independent features; correlated features behave like a smaller number of independent ones, so for smoothed fMRI data it is expected to overestimate the bias, as it does for StarPlus.

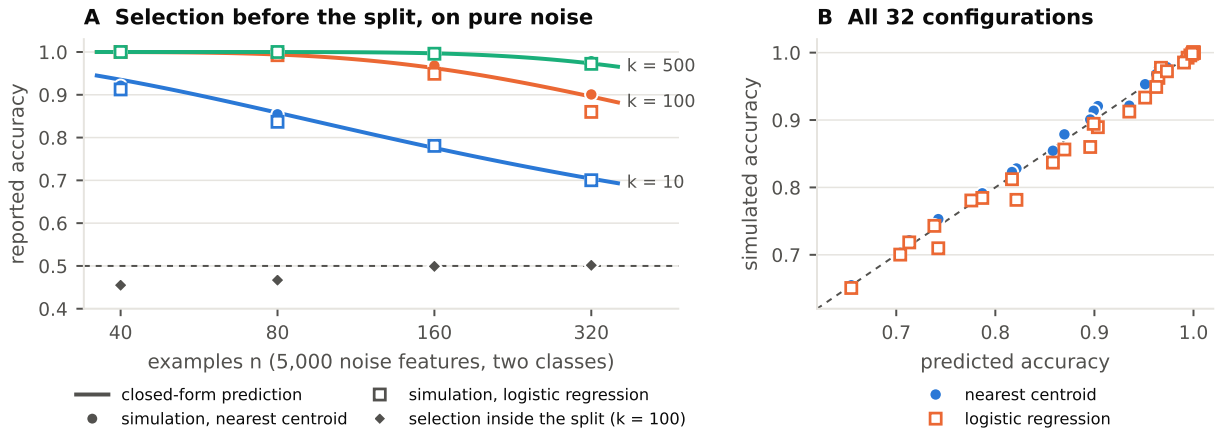

Figure 4: Closed-form prediction of the accuracy reported by selection before the split on pure noise (two classes; chance = 1/2). (A) 5,000 candidate features, $k = 10$, 100 and 500 selected: lines, Eq. (5); filled circles, simulated cross-validation with a nearest-centroid classifier; open squares, logistic regression; diamonds, selection inside each training fold ($k = 100$). (B) Prediction against simulation for all 32 configurations.

E Analysis variants and robustness

An earlier variant of the VT analysis with an 11-fold inner loop and a larger grid gave a mean of 0.717, and a variant that let the inner loop choose between whole-brain and VT candidate pools and offered shrinkage LDA (Ledoit and Wolf, 2004) as a third classifier gave 0.720, choosing the VT pool in 47 of 71 outer folds; neither differs materially from the main analysis. Beta-series features were better for StarPlus (0.929 versus 0.678–0.850 across raw-window variants) but worse for Haxby (Table 10), where the VT restriction helped most for the weakest subjects (subject 2: +19.8 points on beta-series; subject 5: +18.6 on raw frames).

Other controls and classifiers, per subject. Table 11 gives the per-subject accuracies behind Table 4. With C fixed at 1 and a strict subset required, the two label-based rankings differ from each other by −0.4 points on the whole brain and +0.1 points within VT (multi-class ANOVA F minus summed score, mean over subjects), and selection within VT changes logistic-regression accuracy by −1.2 points (90% CI −1.7 to −0.8), compared with −0.4 points in the main analysis, where the inner loop may decline selection. The tuned Gaussian-kernel SVM was fitted on precomputed kernels by a separate implementation; its default-hyperparameter arm reproduces the accuracies of the first implementation in every subject, pool and arm (largest difference 0.0 points).

Negative result: simplified cross-subject alignment. To improve Haxby's weakest subject (subject 4) by pooling other subjects' data (Haxby et al., 2011), within each outer fold we fit an orthogonal Procrustes rotation mapping each auxiliary subject's per-category mean response (in the union of all subjects' VT masks, 1,963 voxels) onto the target subject's training-fold template and pooled the aligned volumes into the training set. Accuracy fell from 0.528 to 0.251. The rotation

Table 10: Haxby per-subject nested accuracy by feature type and candidate pool (WB: whole brain; VT: ventral temporal mask). Beta-series tracks use 96 block patterns per subject; raw-frame tracks 864 volumes (792 for subject 5). The last column is the VT analysis of the main text.

Subject	Beta / WB	Beta / VT	Raw / WB	Raw / VT
1	0.510	0.562	0.745	0.774
2	0.375	0.573	0.619	0.708
3	0.750	0.719	0.735	0.690
4	0.646	0.646	0.593	0.556
5	0.830	0.818	0.591	0.782
6	0.698	0.688	0.784	0.775
Mean	0.635	0.668	0.678	**0.714**

Table 11: Haxby: per-subject nested LORO accuracy for the dimensionality controls and classifiers of Table 4 (mean over outer folds; chance = 0.125).

Classifier	Dimensionality control	S1	S2	S3	S4	S5	S6
Whole brain							
Logistic regression	None	0.424	0.448	0.451	0.394	0.359	0.534
Logistic regression	Summed one-vs-rest F	0.757	0.615	0.735	0.557	0.598	0.793
Logistic regression	Multi-class ANOVA F	0.722	0.631	0.731	0.558	0.592	0.795
Logistic regression	Random voxels	0.343	0.363	0.383	0.329	0.316	0.451
Logistic regression	Highest-variance voxels	0.199	0.204	0.245	0.258	0.268	0.346
Logistic regression	Principal components	0.475	0.441	0.487	0.438	0.395	0.534
Linear SVM	None	0.387	0.381	0.471	0.399	0.367	0.544
Linear SVM	Summed one-vs-rest F	0.758	0.640	0.693	0.579	0.576	0.780
Gaussian-kernel SVM	None	0.138	0.159	0.194	0.157	0.149	0.218
Gaussian-kernel SVM	Summed one-vs-rest F	0.589	0.502	0.639	0.446	0.485	0.634
Gaussian-kernel SVM, tuned	None	0.223	0.293	0.308	0.219	0.230	0.391
Gaussian-kernel SVM, tuned	Summed one-vs-rest F	0.641	0.604	0.699	0.484	0.538	0.763
Prior (VT)							
Logistic regression	None	0.792	0.716	0.672	0.558	0.783	0.763
Logistic regression	Summed one-vs-rest F	0.775	0.711	0.668	0.543	0.768	0.747
Logistic regression	Multi-class ANOVA F	0.782	0.690	0.669	0.554	0.768	0.751
Logistic regression	Random voxels	0.766	0.692	0.669	0.550	0.761	0.751
Logistic regression	Highest-variance voxels	0.777	0.648	0.672	0.559	0.763	0.772
Logistic regression	Principal components	0.770	0.714	0.630	0.590	0.725	0.728
Linear SVM	None	0.796	0.693	0.623	0.579	0.736	0.707
Linear SVM	Summed one-vs-rest F	0.775	0.718	0.615	0.557	0.712	0.700
Gaussian-kernel SVM	None	0.507	0.417	0.578	0.316	0.557	0.581
Gaussian-kernel SVM	Summed one-vs-rest F	0.562	0.476	0.578	0.378	0.557	0.644
Gaussian-kernel SVM, tuned	None	0.726	0.631	0.677	0.472	0.723	0.744
Gaussian-kernel SVM, tuned	Summed one-vs-rest F	0.733	0.645	0.663	0.469	0.741	0.747

was fit from only eight corresponding points in a 1,963-dimensional space, whereas hyperalignment proper uses hundreds to thousands of time points (Haxby et al., 2011; Guntupalli et al., 2016).

Library versions. The VT analysis repeated under `scikit-learn` 1.8.0 / `nilearn` 0.14.0 for 2 subjects changed per-subject accuracies by at most 0.6 points, and the example's four accuracies by at most 0.6 points. The StarPlus ablation reproduced every arm of every subject identically under both version sets. The whole-brain no-selection arm, run twice, gave identical accuracies.