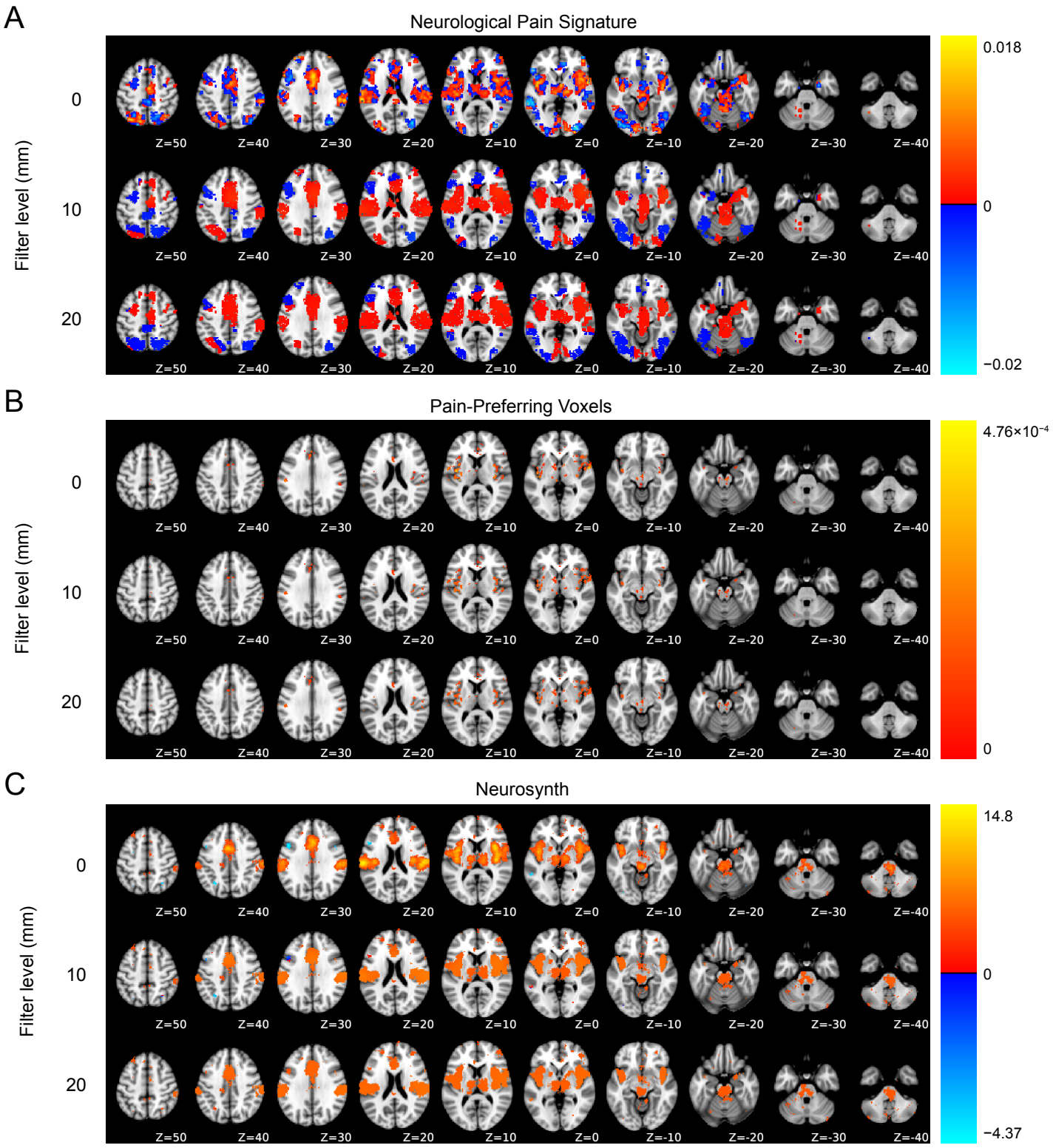
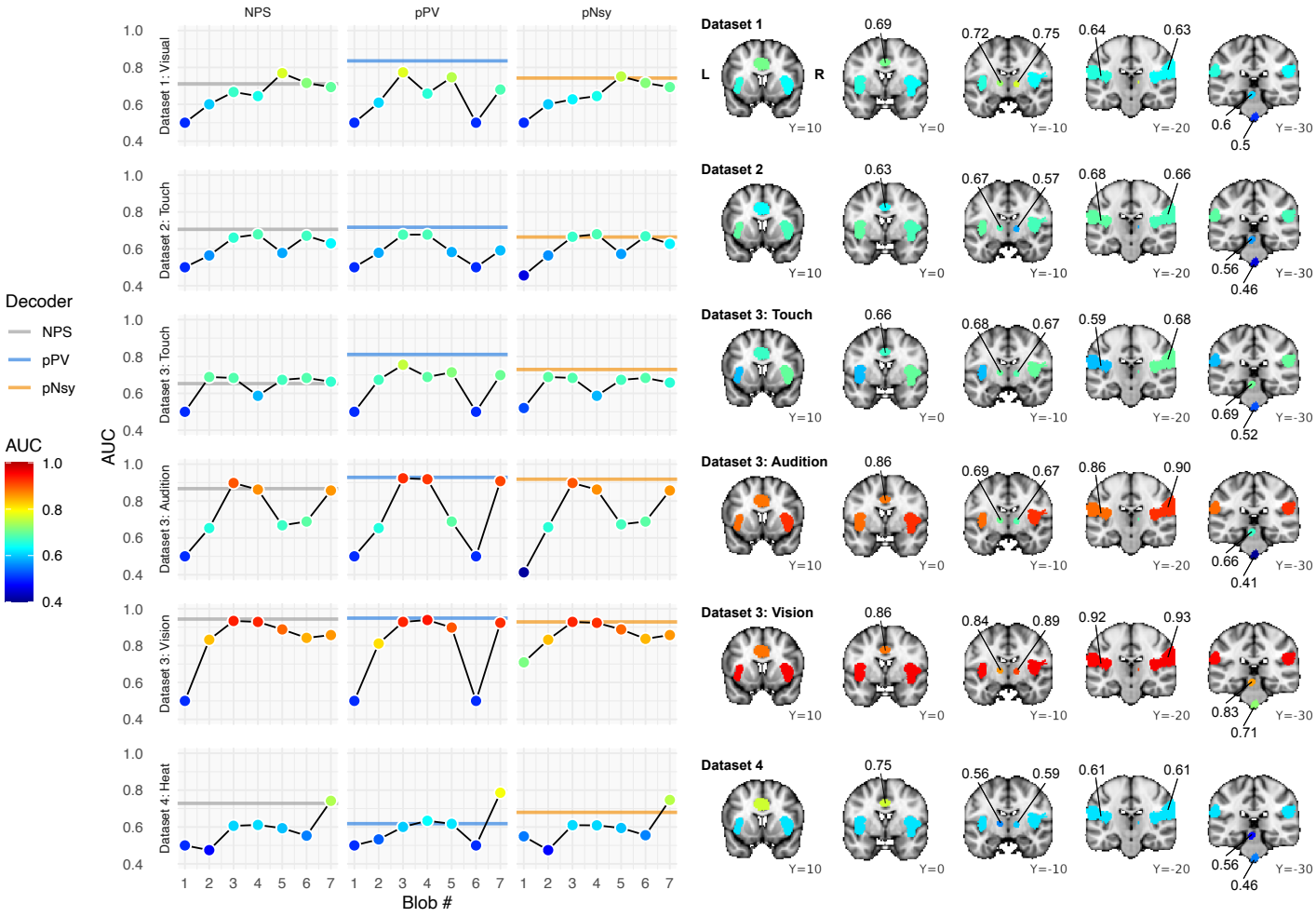
Supplemental Results

Limits of Decoding Mental States with fMRI

R. Jabakhanji
A.D. Vigotsky
J. Bielefeld
L. Huang
M.N. Baliki
G.D. Iannetti
A.V. Apkarian



Supplemental Figure 1. Spatial extents and influence of Gaussian spatial filtering for the three pain decoders. A. The fwMVP Neurological Pain Signature (NPS), **B.** The fwMVP Pain Preferring Voxel (pPV), **C.** The meta-GLM Pain-Neurosynth (pNsy). Presented are dorsal slices from Z=50 to Z=-40 in standard MNI space. Top row in each panel shows the raw, unfiltered decoder. Row 2, shows resultant decoder pattern after 3D spatial smoothing using a Gaussian kernel with a standard deviation of 10 mm (5x voxel width); Row 3, decoder pattern after Gaussian smoothing with a standard deviation of 20 mm (10x voxel width). Note, the differences in spatial patterns and extent between decoders: NPS covers a larger extent of the brain than pPV or pNsy; NPS voxels outside of pNsy are mostly composed of negative values; and pPV is comprised of sparse voxels within the bounds of pNsy. Moreover, the positive weights of all three decoders sample mostly common brain regions (bounded by pNsy). For all three decoders, spatial variability in voxel weights decreases with increasing size spatial filtering: brain regions become more homogeneous; regions with both positive and negative coefficients become either positive or negative with filtering; eventually, at infinity, all template coefficients become a single positive value (at infinite filtering decoders only contain location information: binary decoders: 0, 1). Note, due to the sparsity of pPV, filtering effects are subtle.



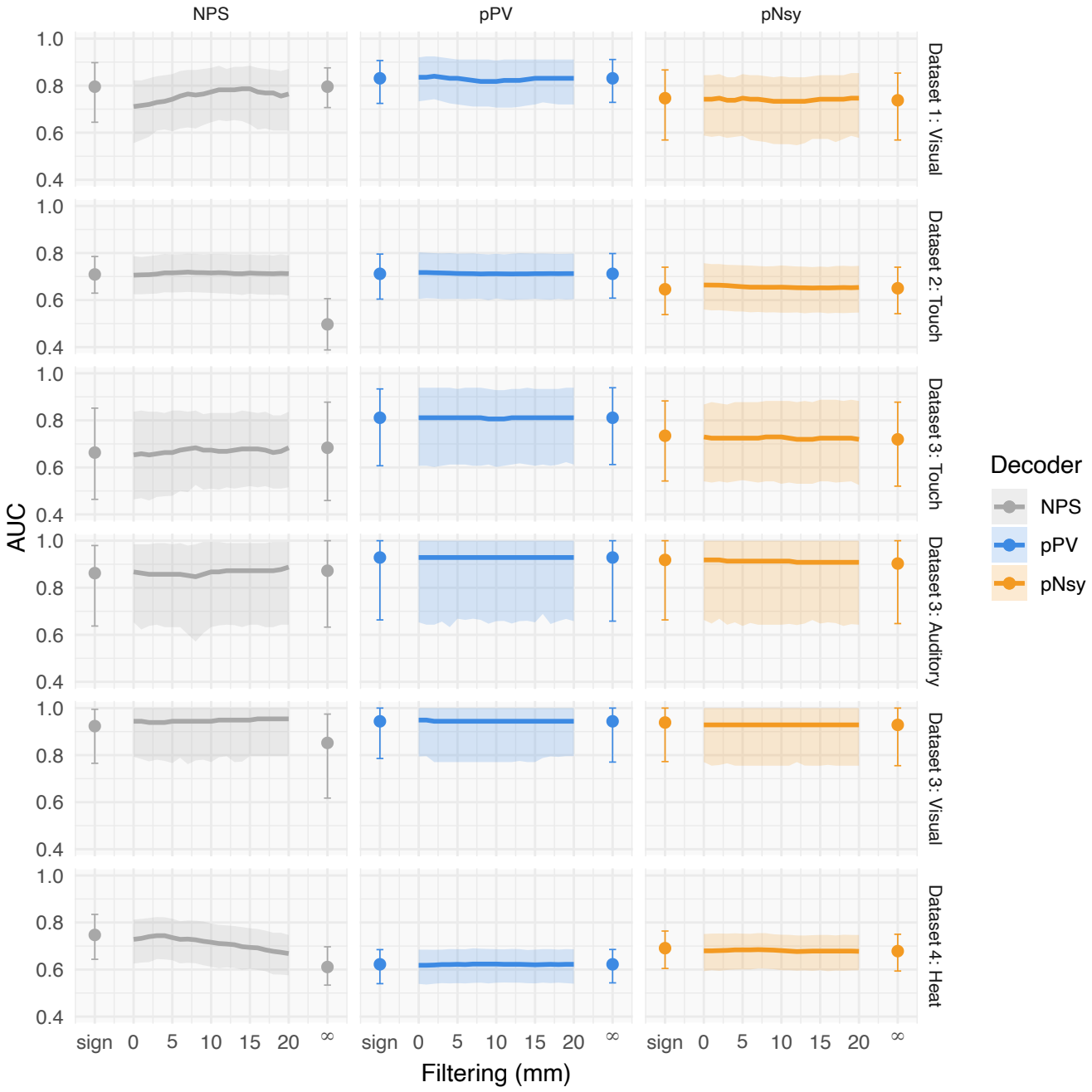
Supplementary figure 2. Discrimination performance for different clusters to assess dependence on specific brain regions. There is a long debate in the literature regarding which brain regions show better specificity to nociceptive stimuli and/or pain states. Here we address the issue from a decoding viewpoint. Here we ask the question whether there are privileged locations for decoding pain. As all three pain decoders (NPS, pPV, and pNsy) overlap most with pNsy, we used the latter to create separate blobs, within each of which we tested the performance of all three decoders. Pain-Neurosynth was thresholded at $z = 6$ to obtain 7 distinct and contiguous regions, or "blobs" (locations illustrated on standard atlas on the right)." NPS, pPV, and pNsy values in each of the 7 blobs were used as separate decoders, and their respective performance was obtained for discriminating pain from comparator states across all studies. Three left columns show discrimination performance for each decoder, within each blob, for all tasks as indicated. The right brain maps show performance for pNsy, localized to each blob (colors match the color distribution in the right pNsy column), for all four task comparisons.

Dataset 1: Visual, painful heat with continuous rating vs visual rating (15 subjects). **Dataset 2: Touch**, painful heat vs tactile stimuli (51 subjects). **Dataset 3: Touch**, Painful heat vs tactile stimuli (14 subjects). **Dataset 3: Auditory**, Painful heat vs auditory stimuli (14 subjects). **Dataset 3: Visual**, Painful heat vs visual stimuli (14 subjects). **Dataset 4: Heat**, Painful heat vs non-painful warm thermal stimuli (33 subjects).

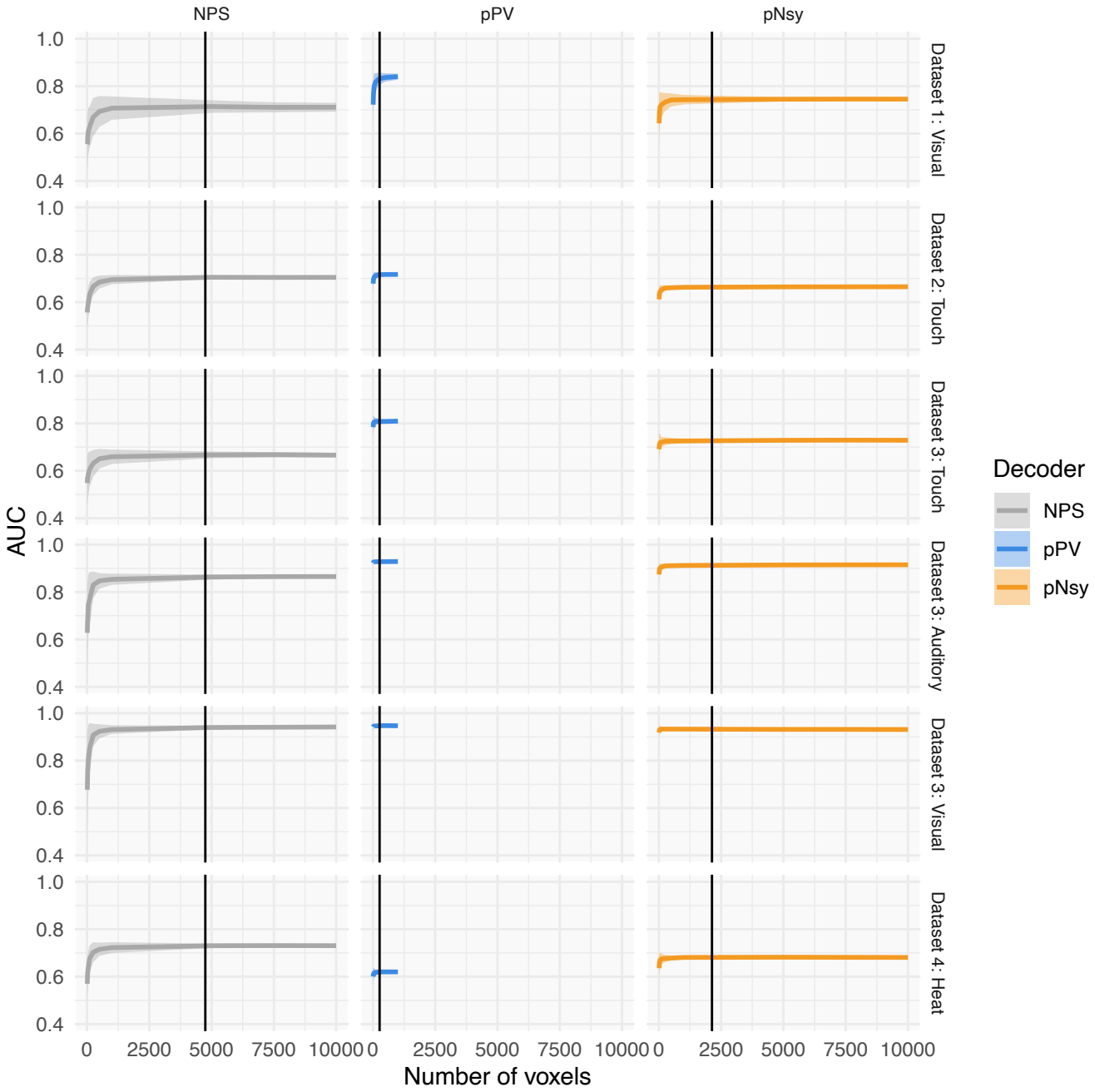
Our performance metric is the area under the receiver operating characteristic curve (AUC).

For any given study, multiple blobs for multiple decoders performed equally well and matched the performance of the full decoder (grey, blue, and orange lines). This result suggests lack of any given blob showing consistently better specificity than others. Some blobs in isolation performed better than the entire decoder (grey, blue, and orange lines), but the effect was study- and decoder- specific. In some instances (Blob #1 with NPS and pPV, Blob #6 with pPV), the Blobs had no spatial overlap with the decoders; for these cases, the blob received an AUC of 0.5. Blob 1 (inferior brainstem) was the region most consistently showing worst decoding ability across studies and decoder types. This is partially explained by the exclusion of Blob #1 in NPS and pPV; but in pNsy, we suspect this effect is due to the influence of physiological noise that contaminates brainstem activity.

Blob 1 = inferior brainstem; blob 2 = superior brainstem; blob 3 = right insula & parietal lobe; blob 4 = left insula & parietal lobe; blob 5 = left thalamus; blob 6 = right thalamus; and blob 7 = anterior cingulate cortex.



Supplemental Figure 3. Discrimination performance as a function of manipulating weights of decoders. We examined relative to raw decoders, preserving sign only (-1, 0, +1) (sign), increasing size spatial smoothing (0-20 mm), and infinite spatial smoothing (0, +1) (∞), for the *original unfiltered* three pain decoders (NPS, pPV, and pNsy), examined in four studies contrasting pain states with various control conditions. **Dataset 1: Visual**, painful heat with continuous rating vs visual rating (15 subjects). **Dataset 2: Touch**, painful heat vs tactile stimuli (51 subjects). **Dataset 3: Touch**, Painful heat vs tactile stimuli (14 subjects). **Dataset 3: Auditory**, Painful heat vs auditory stimuli (14 subjects). **Dataset 3: Visual**, Painful heat vs visual stimuli (14 subjects). **Dataset 4: Heat**, Painful heat vs non-painful warm thermal stimuli (33 subjects).
Our performance metric is the area under the receiver operating characteristic curve (AUC). Values at "sign" are AUC when only the sign of the decoder is preserved (-1, 0, +1); 0 filtering are AUC using raw unfiltered decoders; curves are plots of AUC vs kernel standard deviation for 0-20 mm filtering; ∞ in each graph are performances using filtered templates with infinite standard deviation (binarized templates; 0, +1); whiskers and shaded area are 95% confidence intervals estimated using the bias-corrected and accelerated bootstrap.
Notice that, except for NPS, decoding performance is essentially unaffected by spatial filtering even with an infinitely wide filter which completely deletes the pattern. Generally, these findings suggest that the information captured by all three pain decoders is not dependent on the fine-grained patterns of weights that constitute the decoders.
Performance significantly degrades for infinity-filtered NPS in Dataset 2: Touch (**Row 2**), and slightly for Dataset 4: Heat (**Row 6**). Even in these cases the NPS sign-only shows the same performance as raw NPS, revealing that the degradation is due to assigning positive weights to brain regions where activity is not related to the pain task.



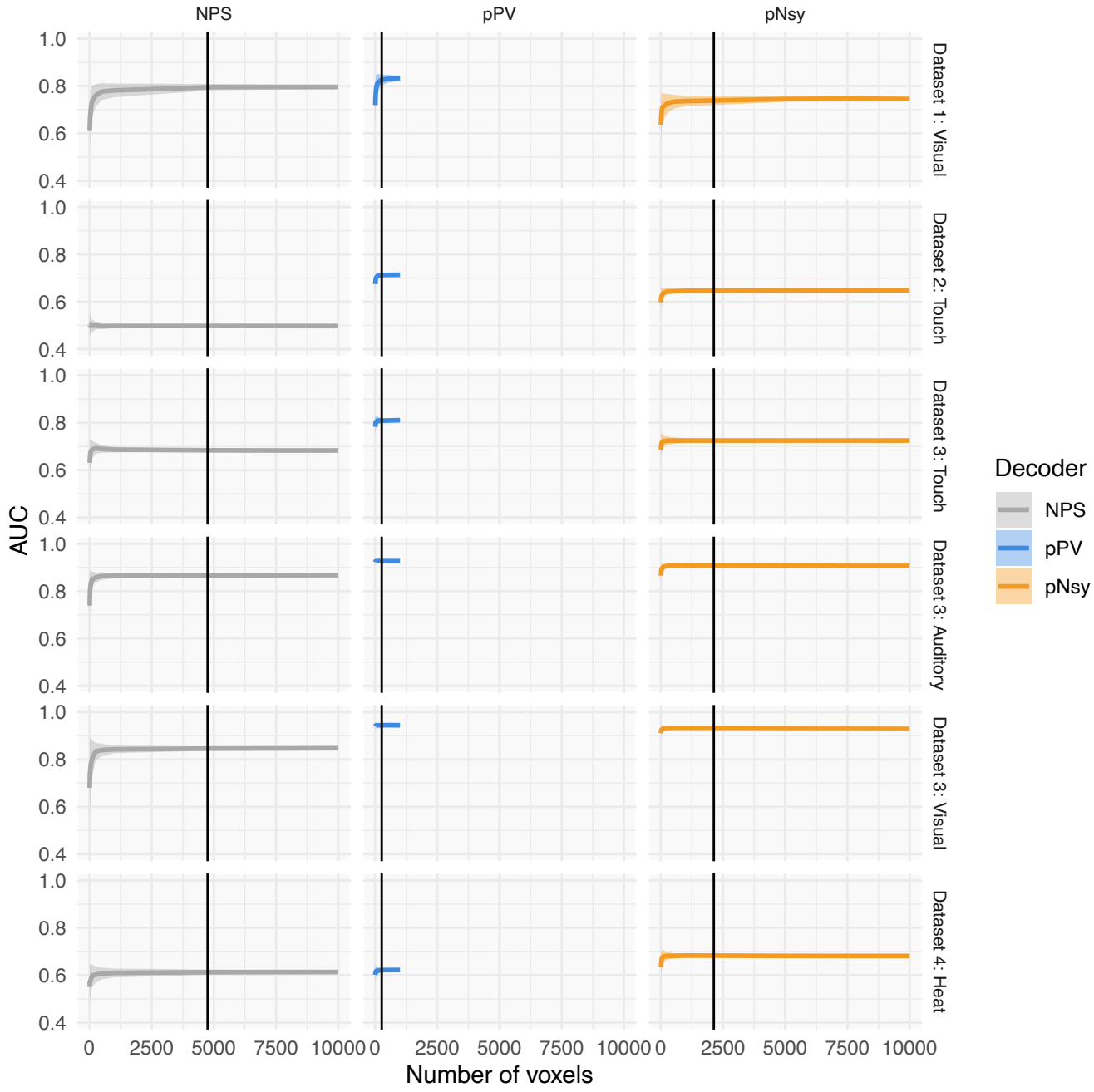
Supplemental Figure 4. Testing for information redundancy by random subsampling of *unfiltered* decoders. For all three *unfiltered* decoders (NPS, pPV, and pNsy; each column), an increasing number of voxels was selected at random to form new decoders. The new decoders were comprised of voxels from 10 up to the total number of voxels in the parent template. Discrimination performance was measured using each new constructed decoder.

Dataset 1: Visual, painful heat with continuous rating vs visual rating (15 subjects). **Dataset 2: Touch**, painful heat vs tactile stimuli (51 subjects). **Dataset 3: Touch**, Painful heat vs tactile stimuli (14 subjects). **Dataset 3: Auditory**, Painful heat vs auditory stimuli (14 subjects). **Dataset 3: Visual**, Painful heat vs visual stimuli (14 subjects). **Dataset 4: Heat**, Painful heat vs non-painful warm thermal stimuli (33 subjects).

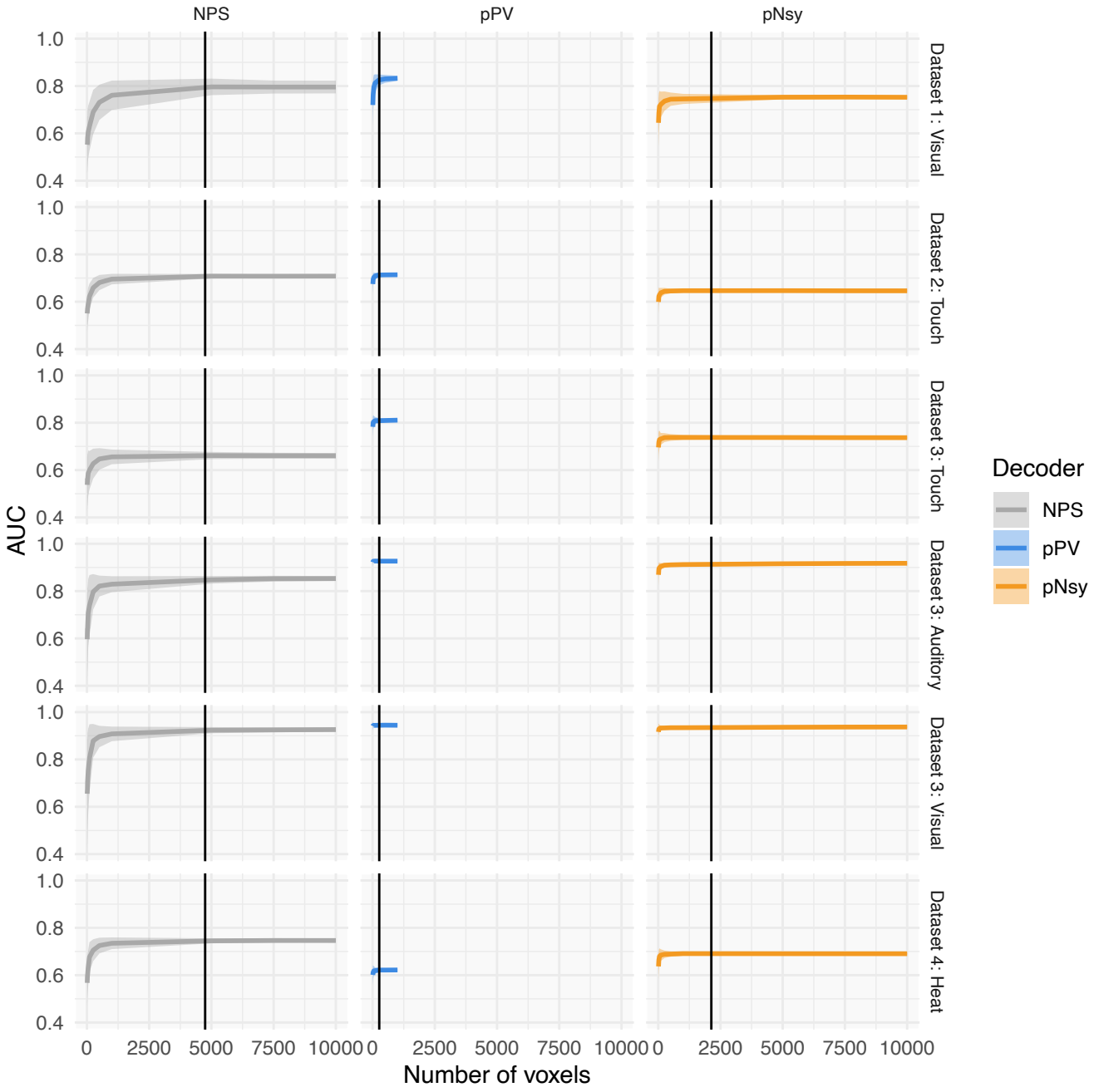
Performance metric is the area under the receiver operating characteristic curve (AUC). Shaded areas are standard deviations due to resampling voxels and thus are insensitive to the number of permutations.

We note that the decoding performance reaches that of its full version very quickly. Less than 10% of the parent decoder was required to reach peak performance (black bar marks).

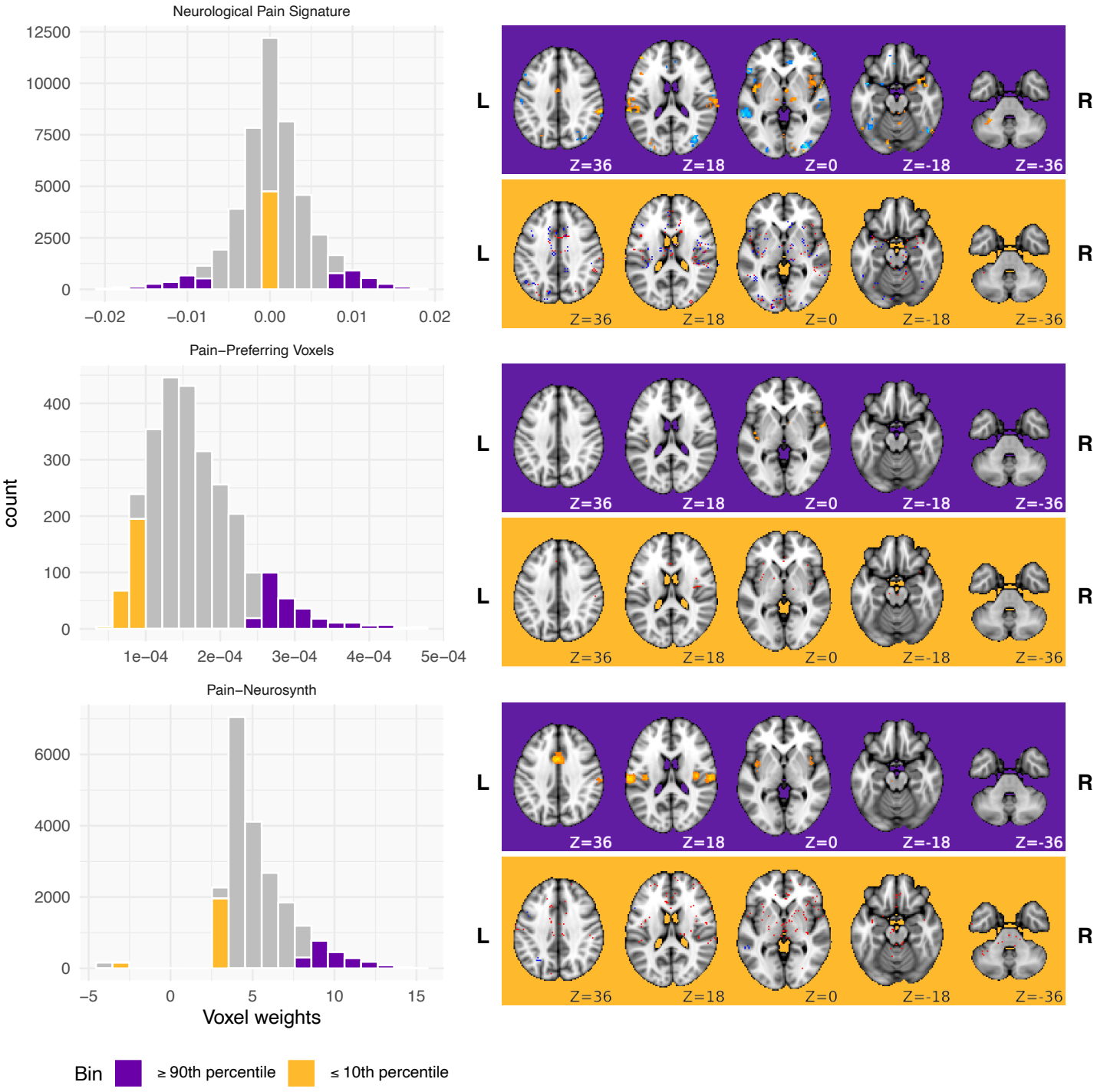
These results indicate existence of sufficient redundancy that on average only a random 10% of the patterns of voxels comprising any of the *unfiltered* decoders was needed to achieve best performance, for all task discriminations.



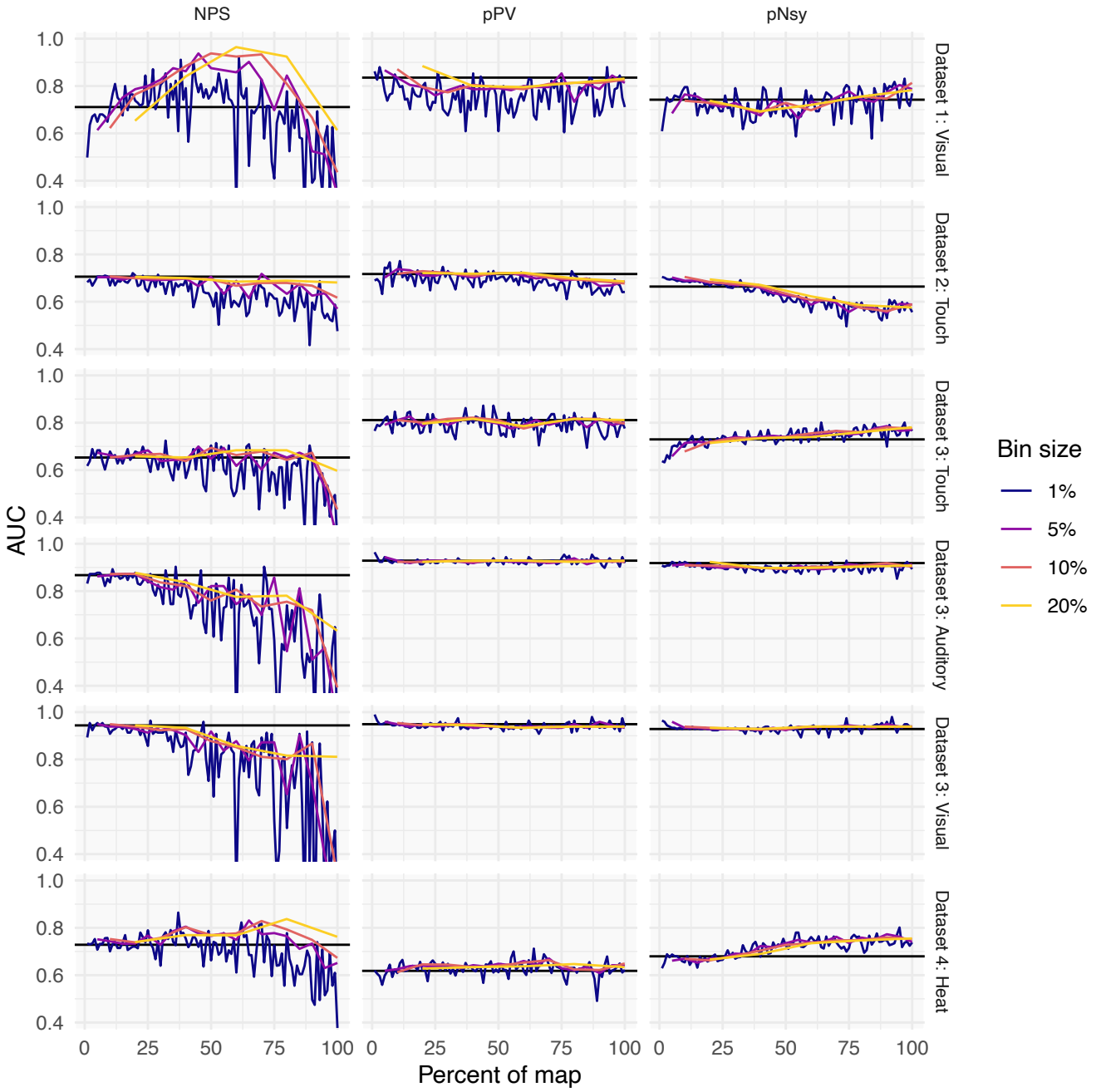
Supplemental Figure 5. Testing for information redundancy by random subsampling of *binarized* decoders. For all three *binarized* (0,1) decoders (NPS, pPV, and pNsy; each column), an increasing number of voxels was selected at random to form new decoders (from 10 voxels to the full template). Classification performance was measured using each new decoder template. Shaded areas are standard deviations due to resampling voxels and thus are insensitive to the number of permutations.
Dataset 1: Visual, painful heat with continuous rating vs visual rating (15 subjects). **Dataset 2: Touch**, painful heat vs tactile stimuli (51 subjects). **Dataset 3: Touch**, Painful heat vs tactile stimuli (14 subjects). **Dataset 3: Auditory**, Painful heat vs auditory stimuli (14 subjects). **Dataset 3: Visual**, Painful heat vs visual stimuli (14 subjects). **Dataset 4: Heat**, Painful heat vs non-painful warm thermal stimuli (33 subjects).
Similar to using raw or unfiltered decoders, we note that the decoding performance reaches peak performance with 10% of the parent decoder (black bars). These results show that only location information of a random 10% of the voxels comprising any of the decoders was sufficient to achieve best performance, in all task discriminations.



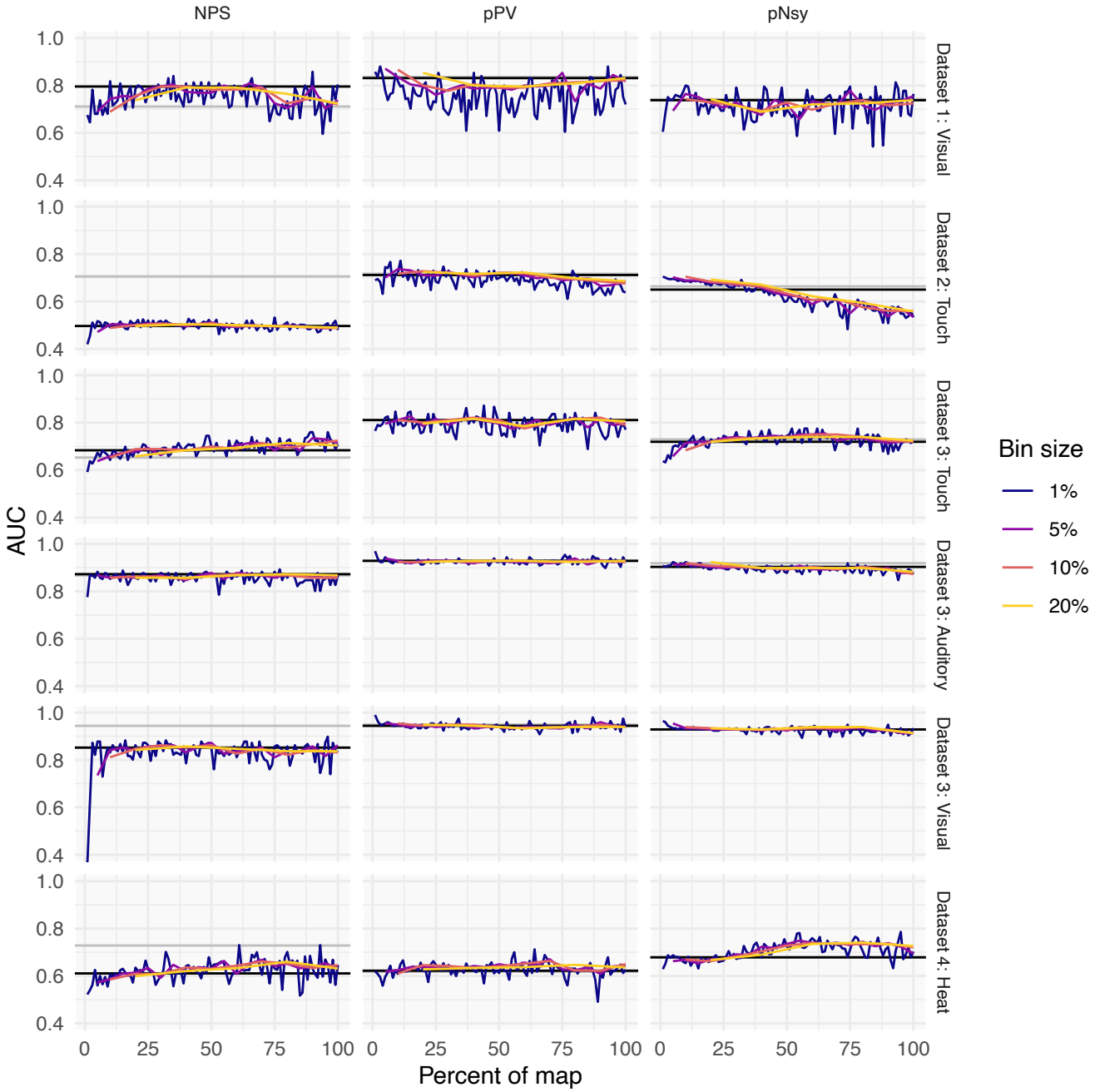
Supplemental Figure 6. Testing for information redundancy by random subsampling of *sign preserving* decoders. For each *sign-preserving* decoder (original templates reduced to three values: -1, 0, +1), an increasing number of voxels was selected at random to form new decoders (from 10 voxels to the full template). Discrimination performance was measured using each new decoder template. Shaded areas are standard deviations due to resampling voxels and thus are insensitive to the number of permutations. **Dataset 1: Visual**, painful heat with continuous rating vs visual rating (15 subjects). **Dataset 2: Touch**, painful heat vs tactile stimuli (51 subjects). **Dataset 3: Touch**, Painful heat vs tactile stimuli (14 subjects). **Dataset 3: Auditory**, Painful heat vs auditory stimuli (14 subjects). **Dataset 3: Visual**, Painful heat vs visual stimuli (14 subjects). **Dataset 4: Heat**, Painful heat vs non-painful warm thermal stimuli (33 subjects).
We again note that the decoding performance reaches peak performance with 10% of the *sign-preserving* decoders (black bars).



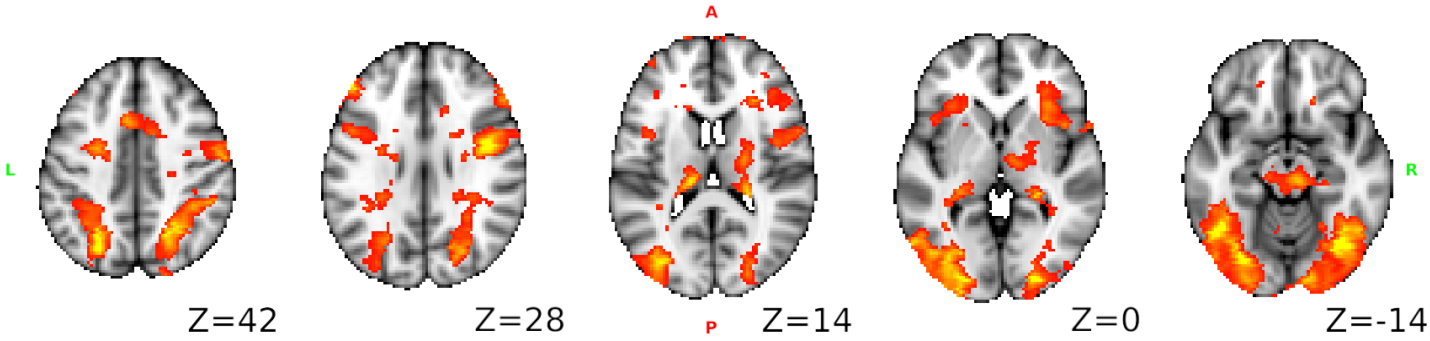
Supplementary Figure 7. Weight-ordered thinning of decoders. The figure illustrates procedure used for weight-ordered thinning. For a given distribution of weights, new decoders were constructed from the absolute value of their coefficients using different bin widths (1%, 5%, 10%, and 20% of the total number of voxels in each decoder). The approach systematically quantifies the contribution of different coefficient-weighted brain regions to discrimination performance. All three pain decoders at least imply the importance of different brain regions based on associated coefficients. In NPS and pPV, brain regions with highest positive and lowest negative coefficients should classify pain better than locations with lower absolute coefficients. In pNsy, weights reflect confidence of discriminating "pain" term associated studies from all other studies included in Neurosynth (association test). Therefore, we used weight ordered thinning to examine differential contribution of the weights to discrimination performance. Results are shown in **Supplementary figures 7-9**.



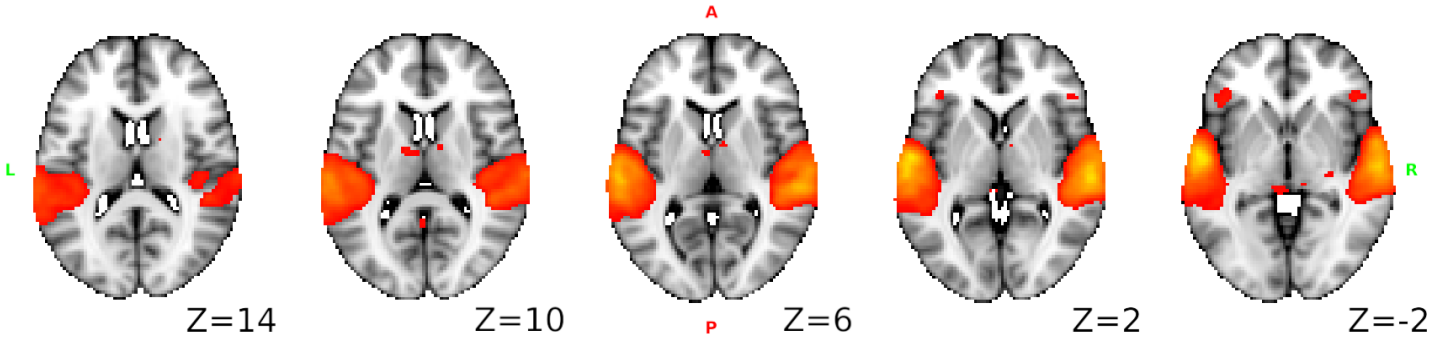
Supplemental Figure 8. Weight-ordered thinning of *unfiltered* decoders. For each decoder (NPS, pPV, pNsy), we binned the voxels according to the absolute value of their coefficients. We used four bin widths: 1%, 5%, 10%, and 20% of the total voxel numbers. We then built a decoding template from each binning, and for each bin width (**Supplementary figure 7**). X-axis of the plot represent the bin number (percent map), Y-axis is performance (AUC), full decoder performance plotted as black line.
Dataset 1: Visual, painful heat with continuous rating vs visual rating (15 subjects). **Dataset 2: Touch**, painful heat vs tactile stimuli (51 subjects). **Dataset 3: Touch**, Painful heat vs tactile stimuli (14 subjects). **Dataset 3: Auditory**, Painful heat vs auditory stimuli (14 subjects). **Dataset 3: Visual**, Painful heat vs visual stimuli (14 subjects). **Dataset 4: Heat**, Painful heat vs non-painful warm thermal stimuli (33 subjects).
If weight distribution was an important determinant for performance then we would expect an inverse relationship between AUC and bin number. Instead we observe mostly an invariant relationship, especially for pPV and pNsy decoders. These findings agree and complement the information redundancy results; demonstrating that the values of voxel coefficients have a minimal and inconsistent impact on decoder performance.
However, the unfiltered NPS shows a weak inverse relationship between AUC and bin number for all datasets and comparator stimuli, most apparently at the smallest bin width (1%) and for bins with absolute values in the lowest 50%. Note when we perform the same analysis now using signed-unfiltered decoders, with binning based on their original weights, the results essentially replicate this figure (data not shown). The latter again refutes, even for NPS, the importance of weight distribution on performance, and instead highlights the fact that in NPS location of negative weights has a small contribution on discrimination performance.



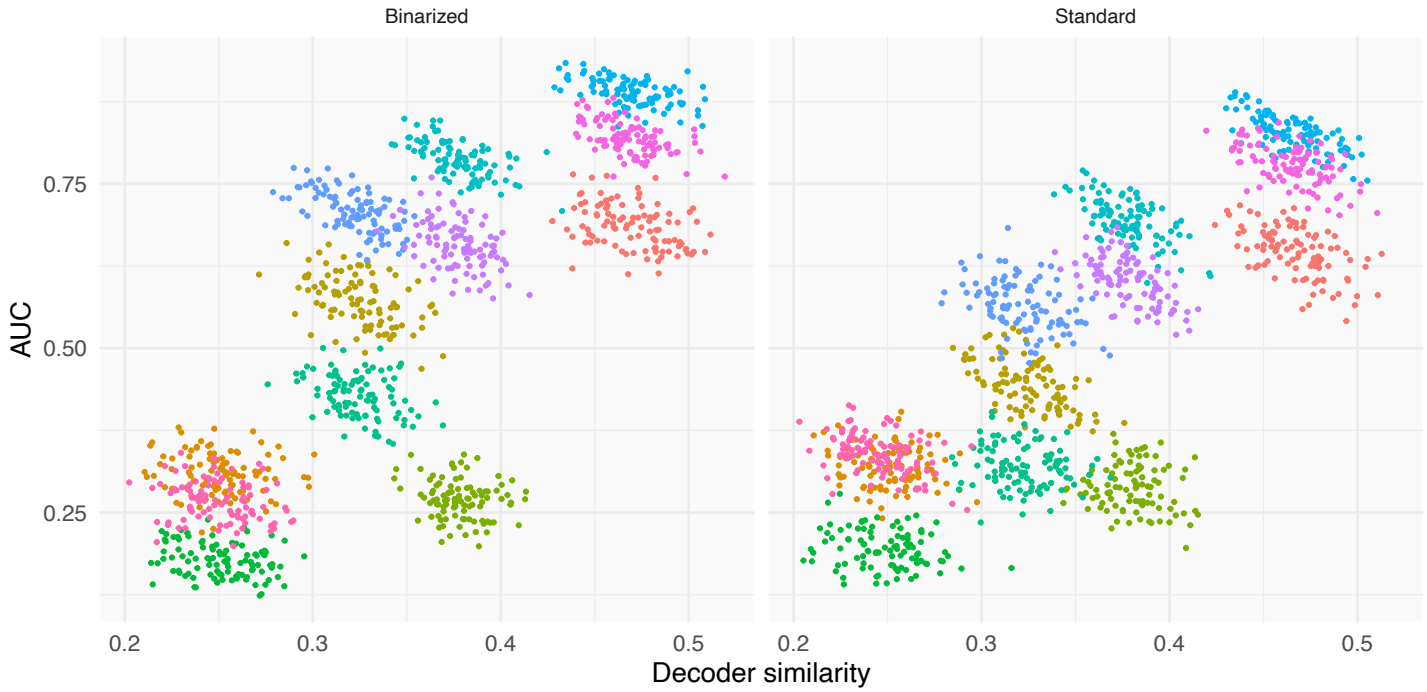
Supplemental Figure 9. Weight-ordered thinning of *binarized* decoders. For each *binarized* (0, +1) decoder (NPS, pPV, pNsy), we binned the voxels according to the absolute value of their original unfiltered coefficients. We used four bin widths: 1%, 5%, 10%, and 20% of the total voxel numbers. We then built a decoding template from the locations for each bin, and for each bin width (**Supplementary figure 7**), but now only using infinitely-filtered decoders (location only: 0, +1). X-axis of the plot represent the bin number (percent map), Y-axis is performance (AUC), full decoder performance plotted as black line.
Dataset 1: Visual, painful heat with continuous rating vs visual rating (15 subjects). **Dataset 2: Touch**, painful heat vs tactile stimuli (51 subjects). **Dataset 3: Touch**, Painful heat vs tactile stimuli (14 subjects). **Dataset 3: Auditory**, Painful heat vs auditory stimuli (14 subjects). **Dataset 3: Visual**, Painful heat vs visual stimuli (14 subjects). **Dataset 4: Heat**, Painful heat vs non-painful warm thermal stimuli (33 subjects).
Strikingly we observe no effect of binning on performance, and at all binnings performance matches that seen for the unfiltered original decoders. Therefore, locations of distinct weightings have no tangible preferential added value to overall performance of these decoders.



Supplemental Figure 10. GLM contrast map for Dataset 5 (Jimura et al.) **(Jimura, Cazalis, Stover, & Poldrack, 2014).** For this dataset, we had access to individual subjects' GLM analyzed activity maps in native space. Thus, the data had to be registered into standard space prior to using it in discrimination and identification analyses, and prior to generating various types of decoders. Therefore, it was imperative to ascertain that the manipulations we had imposed on the dataset still resulted in activity consistent with published results. Using the registered data from Jimura et al., we replicated their results for "mirror-repeat" versus "plain-repeat" task contrast using *randomise* in FSL (z-values, thresholded above 2.3). Our resulting contrast map closely resembles that of Jimura et al. (Jimura et al., 2014).



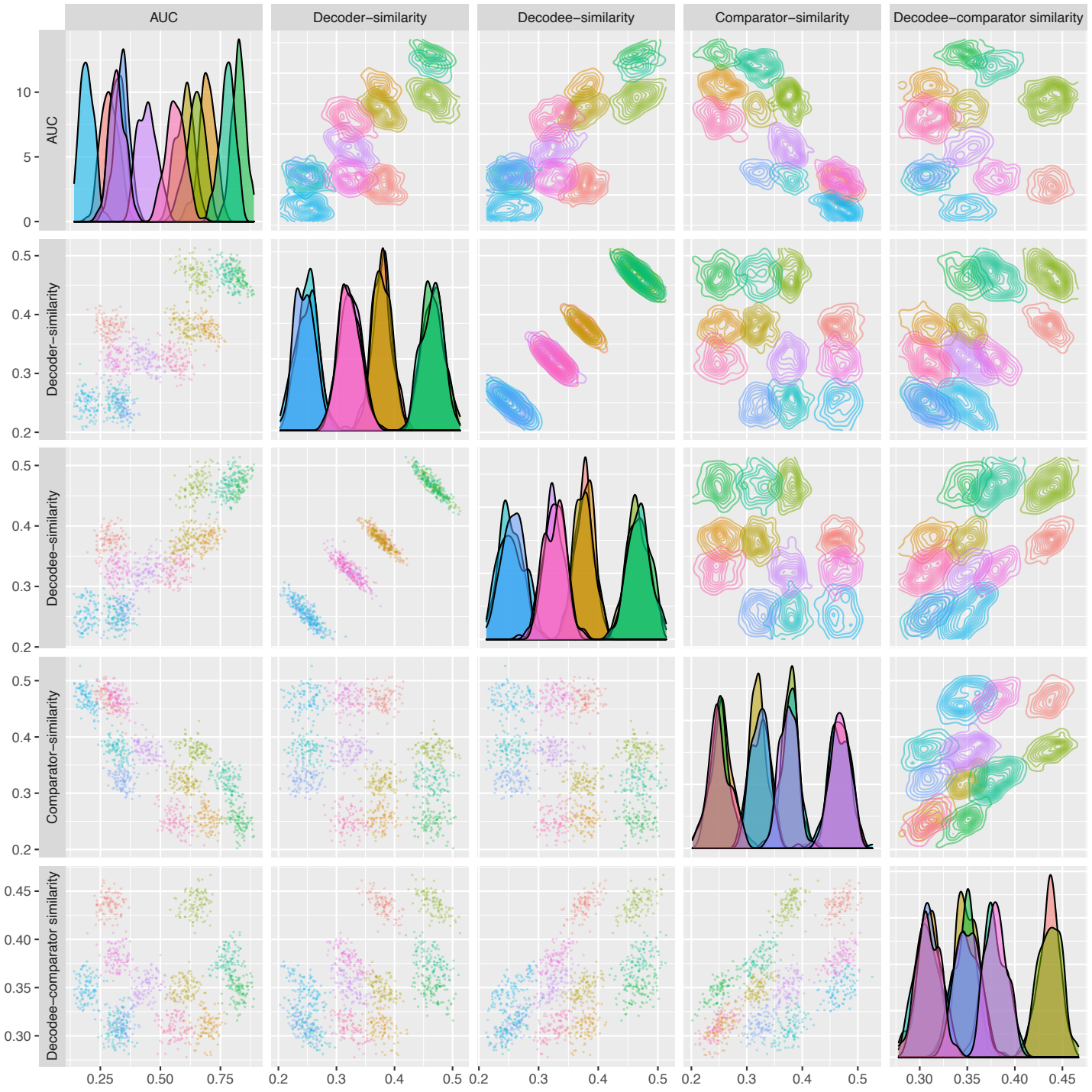
Supplemental Figure 11. GLM contrast maps for Dataset 6 (Pernet et al.) (Pernet et al., 2015). This dataset was only available is raw time series. We, thus, performed standard preprocessing and post-processing to generate individual subject activity maps in standard space. Therefore, it was imperative to ascertain that the manipulations we had imposed on the dataset still resulted in activity consistent with published results. Using the raw data and after our pre- and post-processing, we replicated their results for "vocal" versus "non-vocal" task contrast using *randomise* in FSL (z-values above 2.3). Our resulting GLM contrast map closely resembles that of Pernet et al. (Pernet et al., 2015)



Supplemental Figure 12. Location-only within-subject GLM decoders perform similarly to full GLM decoders. To further assess whether location is the minimum necessary decoding function for successful decoding, we binarized the within-subject decoders from **Figure 4d** as a function of $z > 2.3$.

Left panel is the same as **Figure 6d-left panel**. It is reproduced here to enable direct comparison with location-only within-subject decoders performance.

Within-subject location-only decoders performed similarly to the full decoders regarding the relationship between decoder similarity and AUC. In this case, decoder similarity reflects the extent to which pairs of decoders were spatially coincident. Therefore, even in within-subject decoding, location information is the dominant parameter influencing performance.



Supplemental Figure 13. Determinants of within-subject discrimination performance for GLM decoders. We examined pattern similarity within and between decoder, decodee, and comparator, to uncover determinants of performance. Data from Jimura et al. (Jimura et al., 2014)
were used to create and test within-subject decoders, for four cognitive tasks. Each decoder was created using the averaged GLM activity map from 6 of the 12 task repetitions, and tested on the remaining 6 (decodee) and 6 randomly selected replicates of a control task (comparators). Similarity of the *decoder* represents the average normalized dot product of the GLM maps used to create the decoder, and likewise for the similarity of the *decodee* and *comparator* (i.e., normalized dot products of the GLM maps in each sample). Decodee-comparator similarity is the average normalized dot product between all combinations of the decodee and comparator GLM maps. Note the positive correlations between AUC and decoder and decodee similarities (as shown in **fig. 4D left panel**), and the negative correlation between AUC and comparator similarity; however, there is no salient relationship between AUC and decodee-comparator similarity. Despite this, exploratory regressions found that the best, parsimonious multiple regressions for performance contained both decodee-similarity and decodee-comparator-similarity ($R^2 = 0.95$); the alternative multiple regression model contained comparator-similarity and decodee-comparator-similarity ($R^2 = 0.96$) (see **Supplementary Table 1**). Moreover,

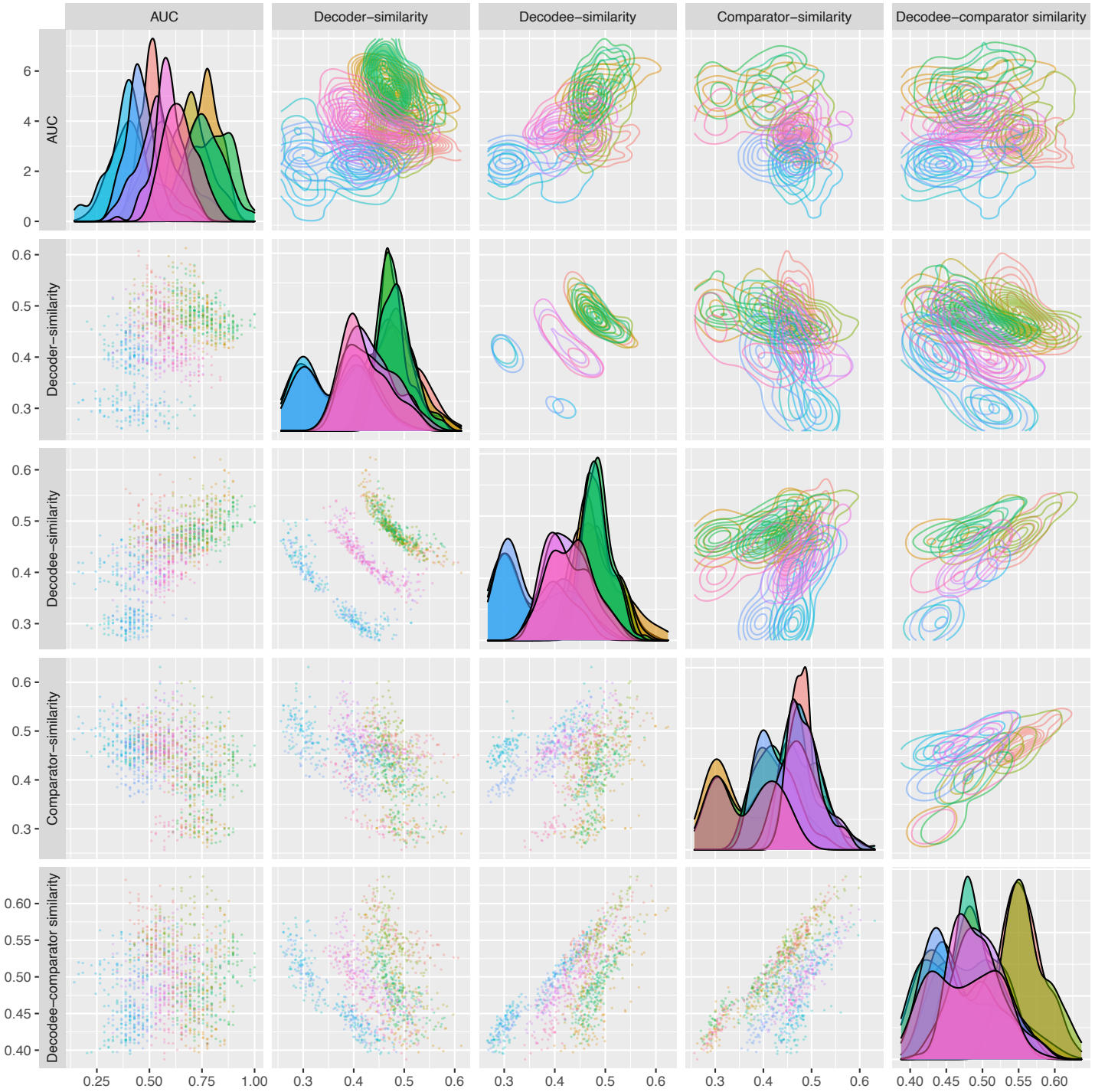
multiple more complex models that are variants of these minimal models also account for most of the variance of decoding. Thus, task properties almost entirely determine within subject decoding performance.
Diagonal panels are density plots, off-diagonal panels are scatter-plots and corresponding contour plots.

Supplementary Table 1. Similarity measures as a determinant of discrimination performance (AUC)

		Within-subject		Between-subject	
Model	**Parameter**	**Estimate ± SE**	**Pseudo R^2**	**Estimate ± SE**	**Pseudo R^2**
1	Intercept	0.02 ± 0.01	0.61	0.38 ± 0.02	0.11
	Decoder similarity	6.69 ± 0.16		2.66 ± 0.21	
2	Intercept	0 ± 0.01	0.66	0.42 ± 0.01	0.38
	Decodee similarity	7.17 ± 0.16		4.83 ± 0.18	
3	Intercept	0.01 ± 0.01	0.58	0.31 ± 0.02	0.11
	Comparator similarity	-6.56 ± 0.17		-2.63 ± 0.22	
4	Intercept	0.01 ± 0.02	0	0.3 ± 0.02	0.01
	Decodee-comparator similarity	0.05 ± 0.42		1.02 ± 0.29	
5	Intercept	0 ± 0.01	0.67	0.44 ± 0.01	0.42
	Decoder similarity	2.06 ± 0.35		1.67 ± 0.18	
	Decodee similarity	5.25 ± 0.36		4.47 ± 0.18	
6	Intercept	0.01 ± 0.01	0.9	0.35 ± 0.02	0.17
	Decoder similarity	5.39 ± 0.09		2.05 ± 0.21	
	Comparator similarity	-5.21 ± 0.09		-2.02 ± 0.22	
7	Intercept	0.02 ± 0.01	0.79	0.34 ± 0.02	0.12
	Decoder similarity	9.11 ± 0.14		2.69 ± 0.21	
	Decodee-comparator similarity	-7.95 ± 0.24		1.16 ± 0.27	
8	Intercept	0 ± 0.01	0.94	0.38 ± 0.01	0.61
	Decodee similarity	5.84 ± 0.07		5.76 ± 0.15	
	Comparator similarity	-5.07 ± 0.07		-4.13 ± 0.15	
9	Intercept	0.01 ± 0.01	0.94	0.84 ± 0.02	0.65
	Decodee similarity	10.71 ± 0.09		9.56 ± 0.21	
	Decodee-comparator similarity	-10.22 ± 0.14		-8.29 ± 0.27	
10	Intercept	-0.01 ± 0.01	0.83	-0.07 ± 0.02	0.38
	Comparator similarity	-9.87 ± 0.14		-7.24 ± 0.27	
	Decodee-comparator similarity	9.69 ± 0.23		7.89 ± 0.34	
11	Intercept	0 ± 0.01	0.94	0.39 ± 0.01	0.61
	Decoder similarity	1.57 ± 0.15		0.22 ± 0.16	
	Decodee similarity	4.38 ± 0.15		5.69 ± 0.15	
	Comparator similarity	-5.04 ± 0.07		-4.04 ± 0.16	
12	Intercept	0.01 ± 0.01	0.94	0.84 ± 0.02	0.65
	Decoder similarity	1.91 ± 0.15		0.46 ± 0.14	
	Decodee similarity	8.92 ± 0.16		9.31 ± 0.22	
	Decodee-comparator similarity	-10.18 ± 0.13		-8.04 ± 0.28	
13	Intercept	0.01 ± 0.01	0.9	-0.04 ± 0.02	0.39
	Decoder similarity	4.64 ± 0.15		0.86 ± 0.2	
	Comparator similarity	-6.01 ± 0.16		-6.72 ± 0.3	
	Decodee-comparator similarity	1.82 ± 0.31		7.44 ± 0.36	
14	Intercept	0 ± 0	0.95	0.74 ± 0.03	0.65
	Decodee similarity	8.32 ± 0.16		8.76 ± 0.28	
	Comparator similarity	-2.63 ± 0.16		-1.17 ± 0.28	
	Decodee-comparator similarity	-5.35 ± 0.32		-6.43 ± 0.53	
15	Intercept	0.01 ± 0	0.95	0.75 ± 0.03	0.65
	Decoder similarity	1.73 ± 0.13		0.33 ± 0.15	
	Decodee similarity	6.83 ± 0.19		8.69 ± 0.28	
	Comparator similarity	-2.48 ± 0.15		-1.01 ± 0.29	
	Decodee-comparator similarity	-5.61 ± 0.3		-6.5 ± 0.53	
16	Intercept	0.01 ± 0.01	0.91	0.36 ± 0.01	0.62
	Decodee / Decodee-comparator	3.68 ± 0.03		4.31 ± 0.10	

We performed beta regression with AUC as the dependent variable and all possible linear combinations of similarity measures as the independent variables. AUCs were shrunken toward 0.5 prior to modeling to force the interval to be (0,1) instead of [0,1] (Smithson & Verkuilen, 2006). All independent variables are mean-centered so that intercepts are interpretable and represent the estimate when all covariates are equal to the mean. Note, standard errors are influenced by the number of permutations and thus should not be used for inference.

In all models (where some of the variance is accounted for; all but model 4), our ability to capture discrimination performance is far superior when comparisons are done within subject, in contrast to between-subject comparisons.



Supplemental Figure 14. Determinants of between-subject discrimination performance for GLM decoders. Data from Jimura et al. (Jimura et al., 2014) were used to create and test between-subject decoders, for four cognitive tasks. Each decoder was created using the averaged GLM activity map from 6 of the 12 task repetitions, and it was tested on 6 other subjects for the corresponding task (decodee-s from different subjects) and 6 other subjects for a comparator task. Similarity of the *decoder* represents the average dot product of the GLM maps used to create the decoder, and likewise for the similarity of the *decodee* and *comparator* (i.e., dot products of the GLM maps in each sample). Decodee-comparator similarity is the average dot product between all combinations of the decodee and comparator GLM maps between different subjects. Note that all relationships are less clear than those in **Supplementary Figure S13**, which define relationships quantified in the regression analysis, shown in **Supplementary Table 1**.

Discussion

Types of Decoders

For the sake of comparison, we distinguish between four different types, although occasionally these categories can be intermixed: 1) *Fixed-weight decoders* (FWD), which is the main topic of our analysis. 2) *Multi-voxel pattern analysis* (MVPA) where within a searchlight window, across-voxel correlational analysis is performed in comparison to conditions or relative to types of stimuli. MVPA can successfully decode diverse brain states from fMRI activity patterns (e.g., (Al-Wasity, Vogt, Vuckovic, & Pollick, 2020; Duff et al., 2015; Haxby, Gobbini, & Nastase, 2020; Haynes & Rees, 2005; Kahnt, 2018; Kaplan & Meyer, 2012; Liang, Su, Mouraux, & Iannetti, 2019; Oosterhof, Tipper, & Downing, 2012; Pilgramm et al., 2016)). 3) Using deep phenotyping or rich naturalistic stimuli, recent work used high-dimensional, multivariable regression stimulus model-based encoders. Such encoders are then transformed into multi-voxel decoders (Naselaris, Kay, Nishimoto, & Gallant, 2011) and are constrained by some dimensionality reduction method (we define these as *naturalistic multi-voxel decoders* (NMVD); for examples see (Cole, Bassett, Power, Braver, & Petersen, 2014; Cole et al., 2013; Dosenbach et al., 2006; Hasson, Nir, Levy, Fuhrmann, & Malach, 2004; Nastase, Gazzola, Hasson, & Keysers, 2019; Varoquaux et al., 2018) and review (Sonkusare, Breakspear, & Guo, 2019)). 4) Decoders of types 1–3 are all based the assumption that information has an anatomically continuous representation in the brain. If one relaxes local contiguity requirements and reorganizes fMRI activity based on functional response similarity, independent of spatial properties, then one can build decoders that outperform type 1–3 decoders. This class of decoders includes diverse approaches. For example, hyperalignment was developed by Haxby and colleagues (Guntupalli et al., 2016; Haxby, Connolly, & Guntupalli, 2014; Haxby, Guntupalli, Nastase, & Feilong, 2020) to efficiently extend MVPA to across-subject decoding using searchlight-based functional alignment, while other methods perform decoding by completely discarding brain anatomy assumptions (C. R. Cox & Rogers, 2021; Kumar, Ellis, O'Connell, Chun, & Turk-Browne, 2020; Rish, 2017) (we define these approaches as *unconstrained multi-voxel decoders* (UMVD)). Below, we briefly review results obtained from type 2–4 decoders to demonstrate that the decoding properties we have uncovered by deconstructing FWD are consistent with and complementary to results observed by 2–4 type decoders.

MVPA distinguishes patterns of multi-voxel activity associated with different stimuli or cognitive states. It is fundamentally used for within-subject identification of locations where fine-grained patterns exist, relative to the task at hand. Thus, the approach makes no assumptions regarding fixed or weighted patterns. MVPA typically uses subject-specific classifier models; as a result, its accuracy drops when predicting other subjects' responses (Al-Wasity et al., 2020; D. D. Cox & Savoy, 2003; Haxby et al., 2011) and can lead to differential performance between within- and across-subject decoding (Clithero, Smith, Carter, & Huettel, 2011). There is strong evidence that spatial smoothing does not degrade MVPA prediction accuracy (Op de Beeck, 2010), implying (consistent with our results for FWD) that fine-grained patterns are not robust and that larger-scale patterns importantly contribute to observed results. It should be emphasized that whatever the source of the MVPA signal, there is very good evidence that the method identifies existences of regional patterns that can be generalized (Avery, Liu, Ingeholm, Gotts, & Martin, 2021; Vermeylen et al., 2020; Wu, Velenosi, & Blankenburg, 2021; Yan, Christophel, Allefeld, & Haynes, 2021), enabling robust novel inferences regarding regional information content (Brooks, Stolier, & Freeman, 2020; Ekstrom, 2021).

NMVD are stimulus model-based encoding and decoding algorithms where brain activity patterns are related to features of stimuli or cognitive states. The obtained decoders generally seem to underlie overlapping and redundant spatial representation, which are generally in close agreement with our results. For example, Chen et al. (2017) (a mixed MVPA and NMVD study) demonstrated brain activity similarities between viewing a movie and a narration from memory. The authors applied searchlight MVPA to identify patterns that could decode scene information from brain activity. Their decoders, much like our data, are coarse in nature and show no apparent patterns at the voxel scale (see Figure 6 from (Chen et al., 2017)). Similarly, Huth, de Heer, Griffiths, Theunissen, and Gallant (2016) (NMVD-type study) performed principal component analysis-based semantic decoding, demonstrating mapping on the spatial scale of gyri (centimeters) rather than voxels (millimeters). Moreover, they show that decoder prediction performance, for each subject and for all subjects, plateaus when 10-20% of features were selected, demonstrating high spatial redundancy. Another NMVD-type study (Hoefle et al., 2018) examined listening to music, where they showed that decoding peaks when only 10-20% of the voxels were used, again implying high spatial redundancy.

Finally, the UMVD approach attempts to enhance decoding by either functionally aligning fMRI data or searching for decoders independent of anatomical considerations (C. R. Cox & Rogers, 2021; Kumar et al., 2020; Rish, 2017). The study by Kumar et al. (Kumar et al., 2020) is most informative as it reveals that visual, auditory, sematic, and object representations are more widespread when data are functionally aligned in comparison to the anatomically-constrained MVPA approach.

References:

Al-Wasity, S., Vogt, S., Vuckovic, A., & Pollick, F. E. (2020). Hyperalignment of motor cortical areas based on motor imagery during action observation. *Sci Rep, 10*(1), 5362. doi: 10.1038/s41598-020-62071-2

Avery, J. A., Liu, A. G., Ingeholm, J. E., Gotts, S. J., & Martin, A. (2021). Viewing images of foods evokes taste quality-specific activity in gustatory insular cortex. *Proc Natl Acad Sci U S A, 118*(2). doi: 10.1073/pnas.2010932118

Brooks, J. A., Stolier, R. M., & Freeman, J. B. (2020). Computational approaches to the neuroscience of social perception. *Soc Cogn Affect Neurosci*. doi: 10.1093/scan/nsaa127

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nat Neurosci, 20*(1), 115-125. doi: 10.1038/nn.4450

Clithero, J. A., Smith, D. V., Carter, R. M., & Huettel, S. A. (2011). Within- and cross-participant classifiers reveal different neural coding of information. *Neuroimage, 56*(2), 699-708. doi: 10.1016/j.neuroimage.2010.03.057

Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S., & Petersen, S. E. (2014). Intrinsic and task-evoked network architectures of the human brain. *Neuron, 83*(1), 238-251. doi: 10.1016/j.neuron.2014.05.014

Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., & Braver, T. S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nat Neurosci, 16*(9), 1348-1355. doi: 10.1038/nn.3470

Cox, C. R., & Rogers, T. T. (2021). Finding Distributed Needles in Neural Haystacks. *J Neurosci, 41*(5), 1019-1032. doi: 10.1523/JNEUROSCI.0904-20.2020

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage, 19*(2 Pt 1), 261-270. doi: 10.1016/s1053-8119(03)00049-1

Dosenbach, N. U., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., . . . Petersen, S. E. (2006). A core system for the implementation of task sets. *Neuron, 50*(5), 799-812. doi: 10.1016/j.neuron.2006.04.031

Duff, E. P., Vennart, W., Wise, R. G., Howard, M. A., Harris, R. E., Lee, M., . . . Smith, S. M. (2015). Learning to identify CNS drug action and efficacy using multistudy fMRI data. *Sci Transl Med, 7*(274), 274ra216. doi: 10.1126/scitranslmed.3008438

Ekstrom, A. D. (2021). Regional variation in neurovascular coupling and why we still lack a Rosetta Stone. *Philos Trans R Soc Lond B Biol Sci, 376*(1815), 20190634. doi: 10.1098/rstb.2019.0634

Guntupalli, J. S., Hanke, M., Halchenko, Y. O., Connolly, A. C., Ramadge, P. J., & Haxby, J. V. (2016). A Model of Representational Spaces in Human Cortex. *Cereb Cortex, 26*(6), 2919-2934. doi: 10.1093/cercor/bhw068

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science, 303*(5664), 1634-1640. doi: 10.1126/science.1089506

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annu Rev Neurosci, 37*, 435-456. doi: 10.1146/annurev-neuro-062012-170325

Haxby, J. V., Gobbini, M. I., & Nastase, S. A. (2020). Naturalistic stimuli reveal a dominant role for agentic action in visual representation. *Neuroimage, 216*, 116561. doi: 10.1016/j.neuroimage.2020.116561

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., . . . Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron, 72*(2), 404-416. doi: 10.1016/j.neuron.2011.08.026

Haxby, J. V., Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2020). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *Elife, 9*. doi: 10.7554/eLife.56601

Haynes, J. D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat Neurosci, 8*(5), 686-691. doi: 10.1038/nn1445

Hoefle, S., Engel, A., Basilio, R., Alluri, V., Toiviainen, P., Cagy, M., & Moll, J. (2018). Identifying musical pieces from fMRI data using encoding and decoding models. *Sci Rep, 8*(1), 2266. doi: 10.1038/s41598-018-20732-3

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature, 532*(7600), 453-458. doi: 10.1038/nature17637

Jimura, K., Cazalis, F., Stover, E. R., & Poldrack, R. A. (2014). The neural basis of task switching changes with skill acquisition. *Front Hum Neurosci, 8*, 339. doi: 10.3389/fnhum.2014.00339

Kahnt, T. (2018). A decade of decoding reward-related fMRI signals and where we go from here. *Neuroimage, 180*(Pt A), 324-333. doi: 10.1016/j.neuroimage.2017.03.067

Kaplan, J. T., & Meyer, K. (2012). Multivariate pattern analysis reveals common neural patterns across individuals during touch observation. *Neuroimage, 60*(1), 204-212. doi: 10.1016/j.neuroimage.2011.12.059

Kumar, S., Ellis, C. T., O'Connell, T. P., Chun, M. M., & Turk-Browne, N. B. (2020). Searching through functional space reveals distributed visual, auditory, and semantic coding in the human brain. *PLoS Comput Biol, 16*(12), e1008457. doi: 10.1371/journal.pcbi.1008457

Liang, M., Su, Q., Mouraux, A., & Iannetti, G. D. (2019). Spatial Patterns of Brain Activity Preferentially Reflecting Transient Pain and Stimulus Intensity. *Cereb Cortex, 29*(5), 2211-2227. doi: 10.1093/cercor/bhz026

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage, 56*(2), 400-410. doi: 10.1016/j.neuroimage.2010.07.073

Nastase, S. A., Gazzola, V., Hasson, U., & Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Soc Cogn Affect Neurosci, 14*(6), 667-685. doi: 10.1093/scan/nsz037

Oosterhof, N. N., Tipper, S. P., & Downing, P. E. (2012). Viewpoint (in)dependence of action representations: an MVPA study. *J Cogn Neurosci, 24*(4), 975-989. doi: 10.1162/jocn_a_00195

Op de Beeck, H. P. (2010). Probing the mysterious underpinnings of multi-voxel fMRI analyses. *Neuroimage, 50*(2), 567-571. doi: 10.1016/j.neuroimage.2009.12.072

Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E., . . . Belin, P. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage, 119*, 164-174. doi: 10.1016/j.neuroimage.2015.06.050

Pilgramm, S., de Haas, B., Helm, F., Zentgraf, K., Stark, R., Munzert, J., & Kruger, B. (2016). Motor imagery of hand actions: Decoding the content of motor imagery from brain activity in frontal and parietal motor areas. *Hum Brain Mapp, 37*(1), 81-93. doi: 10.1002/hbm.23015

Rish, I. C., G.A. (2017). Holographic brain: Distributed versus local activation patterns in fMRI. *IBM J. Res. & Dev., 61*, 3:1-3:9.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods, 11*(1), 54-71. doi: 10.1037/1082-989x.11.1.54

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends Cogn Sci, 23*(8), 699-714. doi: 10.1016/j.tics.2019.05.004

Varoquaux, G., Schwartz, Y., Poldrack, R. A., Gauthier, B., Bzdok, D., Poline, J. B., & Thirion, B. (2018). Atlases of cognition with large-scale human brain mapping. *PLoS Comput Biol, 14*(11), e1006565. doi: 10.1371/journal.pcbi.1006565

Vermeylen, L., Wisniewski, D., Gonzalez-Garcia, C., Hoofs, V., Notebaert, W., & Braem, S. (2020). Shared Neural Representations of Cognitive Conflict and Negative Affect in the Medial Frontal Cortex. *J Neurosci, 40*(45), 8715-8725. doi: 10.1523/JNEUROSCI.1744-20.2020

Wu, Y. H., Velenosi, L. A., & Blankenburg, F. (2021). Response modality-dependent categorical choice representations for vibrotactile comparisons. *Neuroimage, 226*, 117592. doi: 10.1016/j.neuroimage.2020.117592

Yan, C., Christophel, T. B., Allefeld, C., & Haynes, J. D. (2021). Decoding verbal working memory representations of Chinese characters from Broca's area. *Neuroimage, 226*, 117595. doi: 10.1016/j.neuroimage.2020.117595