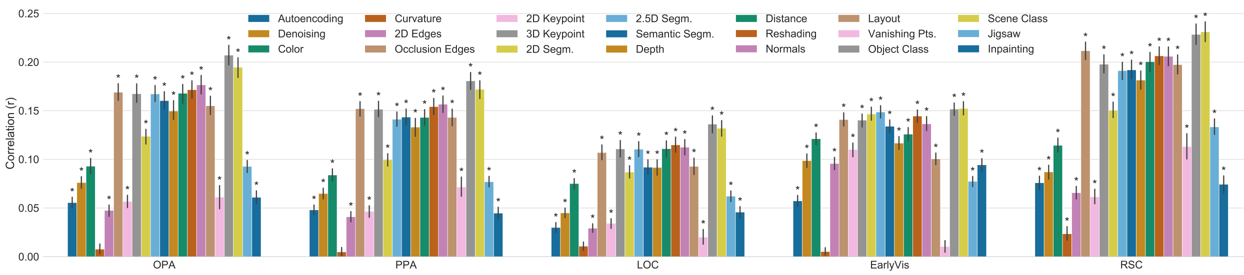
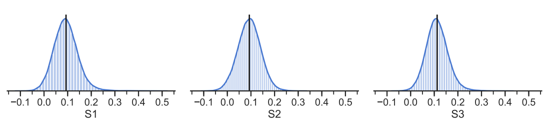
We thank the reviewers for their useful feedback.

R1&R3, Accuracy baseline and ceiling: P-values under the assumption of independence can be problematic, but here: (1) We have fMRI events that are separated by 10 seconds. (2) Before training, we z-scored within sessions and runs. We trained and predicted with shuffled, unordered single trials, but not a time series of data. (3) The BH FDR controls the false discovery rate under positive dependence which is a very common assumption for fMRI. To further minimize assumptions on the underlying distributions, we ran permutation test where we shuffled responses 5000 times, computed the correlation scores, and obtained FDR corrected p-values for both ROI results (shown in figure below for Subject 1; * = FDR corrected $p < 0.00001$) and whole brain results.



Across the whole brain, p-values obtained through permutation tests show more significant voxels across tasks. To provide a better estimate of the variance ceiling, we ran ridge regression to predict between subjects. In the figure below we show the prediction correlation for each subject from other two subjects (each subfigure is a histogram of correlation scores across voxels). The average correlation between predictions and true responses across voxels for each subject are: 0.0931, 0.0932, 0.112, as shown by the black lines on each plot (this includes low SNR voxels that are not engaged by the task). The accuracy we obtained on the significant voxels across the task is close to the ceiling.



R1, Other Taskonomy tasks: Out of the 25 tasks that are provided by the Taskonomy pre-trained task bank, 4 of them take multiple images as input and therefore are excluded in this analysis since brain responses are to single images. For the 21 single image tasks, 2 of them (jigsaw and inpainting) were not included in the paper since they are less human related. Nevertheless, the features from these tasks still encode important information and significantly predict both whole brain voxels and ROIs, as shown in first figure above. Because of limited space, we do not include the whole brain prediction map.

R1-3, To quantify the consistency of results across subjects, we computed correlations of prediction accuracy across subjects: 0.7957 (S1 vs. S2), 0.9034 (S1 vs. S3) and 0.9345 (S2 vs. S3). It would be more difficult to quantify consistency outside of ROIs since structurally projecting one whole brain onto another blurs functional data. We also computed correlations of task distance matrices across subjects: 0.9128 (S1 vs. S2), 0.9304 (S1 vs. S3) and 0.8884 (S2 vs S3). Tree structures of tasks can be obtained through hierarchical clustering on these pairwise distance matrices. Therefore the high correlations across subjects reinforce the **stability of task tree structures (R1)**. **(R2&R3)** Task similarity does not necessarily imply similar predictions in the brain. It depends on the degree of overlap between tasks - if they have very high overlap they will predict basically the same brain data. But if tasks overlap less (say 50% of the variance), then there is plenty of "room" for each task to predict unique brain areas based on the non-shared variance.

R3&R1 For all encoding models, we predict an average of TR3 and TR4. Regression regularization hyperparameters are fit for individual voxels and individual subjects. To avoid overfitting, only validation data but not test data is used to pick the hyperparameters. The task similarity matrix presented in Figure 5 in our paper is an average of distances from 3 subjects (using cosine similarity). Curvature and autoencoding have < 1 similarity to themselves because for some subjects they predict zero significant voxels (zero vectors is undefined for cosine distance, and 0 was returned). We also used other similarity metrics and the patterns were very similar. **(R1)** Prediction maps in the whole brain do predict novel areas beyond the pre-defined ROIs. However, making claims about new functionally-defined brain areas would be premature given our current data and analyses - to the detriment of the field, such claims are often made based on inconclusive data (so-called "fishing expeditions"). To make a robust claim about new "functional territories" we expect to first run additional validation experiments in which specific, theory-driven manipulations were used to establish that specific brain regions are sensitive to the task in question.