

Supplementary material

Using expression-masked images

Images in the Allen dataset are provided in two formats: the raw imagery, and images that were processed as previously described¹ to remove the background, yielding expression-masked images. The analysis was applied to the masked images. This is a big advantage when examining expression patterns, as noise effects coming from cytoarchitecture and underlying brain structures is reduced. Examples of a pair of images are given below in Fig S1.

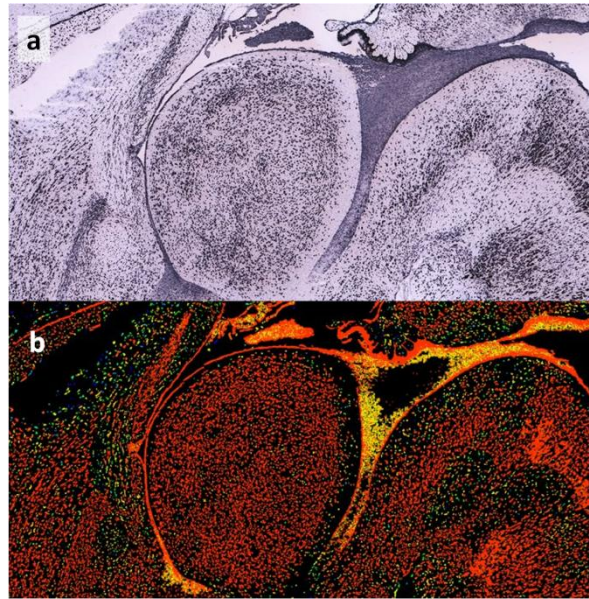


Figure S1: Regular (a) and expression-masked (b) examples of ISH images as provided by the Allen Brain Atlas, for the gene *Tuba1*. While the expression masked images are presented in color, the color images are in fact derived from gray-scale images, which we have used in this work.

Robustness of bag-of-words representations

In order to validate the stability of the bag-of-words gene representations, we measured the similarities between pairs of representations of images that are of the same gene but from different image series, and the similarities between the representations of different genes.

Similarity is much higher for representations of the same gene (Wilcoxon difference of medians test, $p < 10^{-200}$). The similarity values are shown in figure S2. This implies that representations of the same gene, derived from different image series are indeed stable and are representative of the gene.

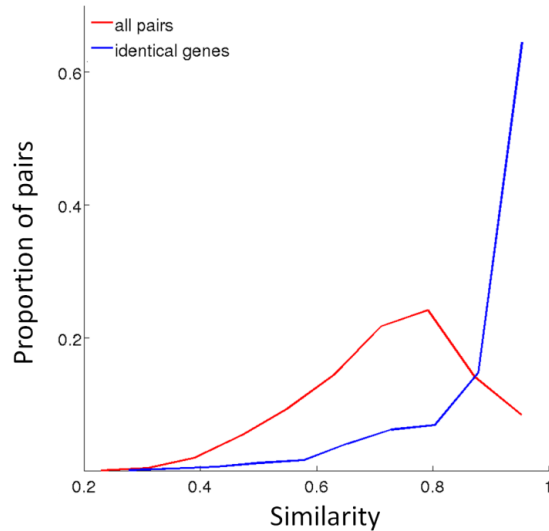


Figure S2: The similarity in the representation of same-gene pairs (blue) and different-gene pairs (red). Each curve shows the histogram of similarity values. Same-gene image series have highly similar representations.

Choosing the dictionary size

In order to choose the size of the visual word dictionary, we performed analysis with four dictionary sizes: 100, 200, 500 and 1000. Figure S3 shows mean test-set AUC values obtained using the different dictionary sizes. Mean AUC across categories is insensitive to the size of the dictionary (K). To check how stable the representations are between the different K 's, we measured the Pearson correlation between AUC values of the 2081 GO categories using the different dictionary sizes. Correlation values are very high and are shown in table R1. The lowest correlation value is 0.846, between $K=100$ and $K=1000$, and is still highly significant ($P < 10^{-100}$). Correspondence between AUC values for the 2081 GO categories obtained using the two dictionary sizes are shown in figure S4, showing indeed a high linear correspondence.

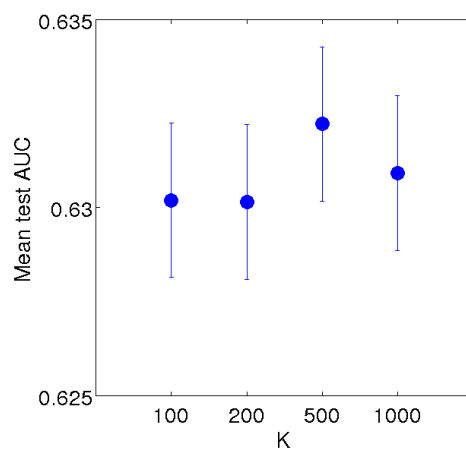


Figure S3: Mean test-AUC values for dictionary size $K=100, 200, 500, 1000$. Error bars indicate standard error of mean across five folds in cross-validation data.

Dictionary size (K)	100	200	500	1000
100	1	0.896	0.861	0.846
200	0.896	1	0.896	0.883
500	0.861	0.896	1	0.917
1000	0.846	0.883	0.917	1

Table S1: Pearson's rho correlation values between AUC results for 2081 categories, compared across the 4 different dictionary sizes. Correlations are high (the lowest is 0.846 between K=100 and K=1000)

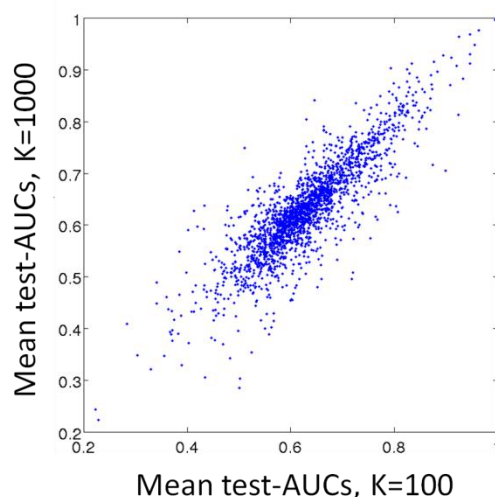


Figure S4: Mean test-set AUCs for dictionary size K=100 versus K=1000. This pair of dictionary sizes is the least correlated among all dictionary size pairs. It can be seen that even in this case, the correlation is high and indicative of a stable representation.

Choice of GO category size:

We chose GO categories with a number of annotations ranging from 15 to 500 genes. We set the lower limit to 15 in order to provide enough positive examples for testing the classifiers across five cross-validation partitions. The higher limit is set to 500 to preclude the resulting semantic explanations from being very general (we use more specific categories such as "*regulation of long-term neuronal synaptic plasticity*" or "*glutamate receptor signaling pathway*" and avoid general categories such as "*transport*" or "*biological regulation*").

To make sure that this choice of categories did not cause a bias in the classification results, we checked the relation between category size and test-set AUC scores. No significant relation between the size of the GO category and the resulting AUC values (Figure S5).

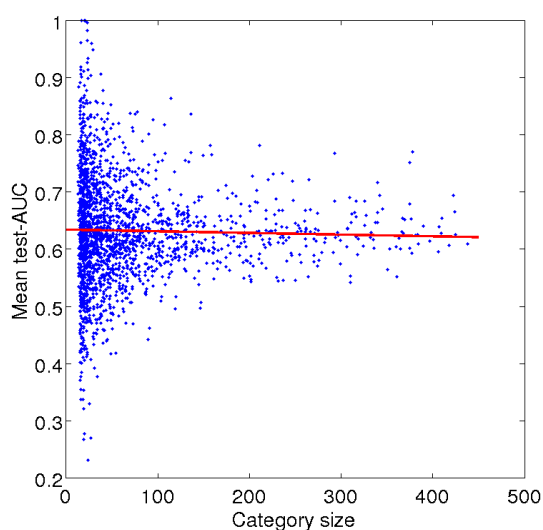


Figure S5: Mean AUC (averaged over test-splits) for the GO categories vs. GO category size (number of genes in the category). There's no significant relation between classification success of a category and the number of genes annotated to it.

Using several slices from each image series

In order to take into fuller account the 3D structure of the brain, we repeated the full set of our experiments while including two additional sagittal sections. The three sections used were taken from one hemisphere, capturing the medial section and also the 30% and 50% marks on the medial-lateral axis. An example of three such slices is shown in Figure S6.



Figure S6: Each image series was represented with three slices, the most medial (a), and the 30% (b) and 50% (c) marks on the medial-lateral axis.

The results of the experiments using multiple slices were inconclusive. In some measures of performance, such as the correlation of our funcISH scores with known PPI interactions, adding more slices has improved the correlations. In others, such as correlations with cell types and pathways, the performance measures did not improve and even deteriorated slightly. The reasons for this inconsistency could be that the location of the non-medial slices is more variable, due to variation across

brains. We note that in the main paper we report the results using a single medial slice.

Applying a spatial pyramid kernel to the images

A major goal of brain-image analysis is to develop a representation that captures both low level texture and gross-anatomy structure. While the visual bag-of-words representation we have used in our work removes global structures, a main advantage is the ability to find small-scaled spatial patterns that are location-independent in the brain.

To combine local patterns with global structures in the same representation, we tested a representation of the data using spatial pyramid kernels². In this approach, every image is split into 4 and 16 rectangles and the bag of words method is applied to each rectangle separately (Figure S7). The resulting feature vector is a concatenation of the $1+4+16 = 21$ dictionaries. This approach has been shown to be highly successful in machine vision tasks^{3,4}. The down side of this approach is that it inflates the feature dimensionality significantly, and requires reducing the dictionary size. In our experiments, we tested a dictionary size 100, which provides similar accuracies as the dictionary size of 500 used in the rest of the analysis (as shown above).

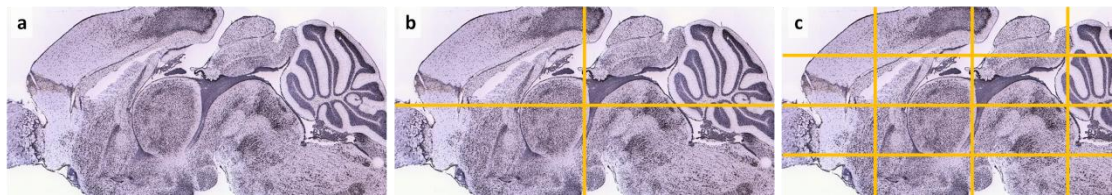


Figure S7: A spatial pyramid approach to extracting dense SIFT features. Features were extracted in the full image (a) and the image divided into four parts (b) and 16 parts (c).

The spatial pyramid approach yielded an overall mean AUC of 0.6231, which is slightly and insignificantly lower than the mean AUC obtained without the pyramidal kernel, 0.6322. We conclude that the increase in feature dimensionality hurts more than the gain obtained by describing different brain regions separately.

These results illustrate the challenging tradeoff when computing both local and global features. An alternative approach could be based on data-dependent segmentation of images into anatomic structures (like the thalamus, cortex or cerebellum) followed by coding each structure separately. Such segmentation is a topic for a separate research.

1. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–76 (2007).

2. Lazebnik, S., Schmid, C. & Ponce, J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Volume 2 CVPR06 2*, 2169–2178 (2006).
3. Grauman, K. & Darrell, T. *The pyramid match kernel: discriminative classification with sets of image features*. *Tenth IEEE International Conference on Computer Vision ICCV05 Volume 1 2*, 1458–1465 (IEEE: 2005).
4. Huang, T. Linear spatial pyramid matching using sparse coding for image classification. *2009 IEEE Conference on Computer Vision and Pattern Recognition* 1794–1801 (2009).doi:10.1109/CVPR.2009.5206757