Improved Unsupervised Representation Learning of Spatial Transcriptomic Data with Sparse Filtering

Mohammad Abbasi, Connor R Sanderford, Narendian Raghu, Benjamin B Bartelle
Arizona State University

ABSTRACT:

We have developed representation learning methods, specifically to address the constraints and advantages of complex spatial data. Sparse filtering (SFt), uses principles of sparsity and mutual information to build representations from both global and local features from a minimal list of samples. Critically, the samples that comprise each representation are listed and ranked by informativeness. We used the Allen Mouse Brain Atlas gene expression data for prototyping and established performance metrics based on representation accuracy to labeled anatomy. SFt, implemented with the PyTorch machine learning libraries for Python, returned the most accurate reconstruction of anatomical ground truth of any method tested. SFt generated gene lists could be further compressed, retaining 95% of informativeness with only 580 genes. Finally, we build classifiers capable of parsing anatomy with >95% accuracy using only 10 derived genes. Sparse learning is a powerful, but underexplored means to derive biologically meaningful representations from complex datasets and a quantitative basis for compressed sensing of classifiable phenomena. SFt should be considered as an alternative to PCA or manifold learning for any high dimensional dataset and the basis for future spatial learning algorithms.

INTRODUCTION:

Representation learning methods have become an essential complement to omics scale studies. Classical dimensionality reduction techniques like PCA, manifold learning methods like tSNE, and clustering methods like k-means all generate more tractable representations (components, features, dictionaries, clusters, etc.) from high dimensional data that would be otherwise uninterpretable to a human mind. Ideally, a representation produces intelligible, even actionable, information, however their utility is inconsistent. Many only present noise or render underlying biological data unintelligible. We sought to develop an optimal method for generating lower dimensional representations with maximal biological significance, while retaining actionable information about the inputs. Essentially, we asked if unsupervised representation learning could generate functional or even predictive models from a biological dataset with minimal implicit priors.

The primary purpose of a representation has been to present complexity as a simple visual plot to support a narrative (Chari et al., 2021). Spatially resolved data presents a challenge for established learning methods, wherein spatial and biological

data are convolved. Data from functional molecular imaging or spatial transcriptomics loses critical information when projected into feature space, demanding a different approach. The Allen Mouse Brain Atlas (AMBA) offers a prime example, wherein expression ~60 voxels of energy data for ~20k genes are arranged within a 67 × 41 × 58 matrix of otherwise blank voxels, labeled with anatomy established by thousands of person hours worth of labor (Ng et al., 2009). We chose AMBA as a prototyping dataset because using gene expression data to predict anatomical "ground truth," has been well explored.

The rapid advance of machine learning (ML) methods, like representation learning, has demonstrated the power of biologically inspired principles of information theory. While almost all ML is by definition emulation of biological phenomena, we hypothesized that a learning method based on the principles of biological networks offered the best strategy for extracting meaningful representations from biological data. Two methods in particular both describe biological systems and have demonstrated effectiveness in representation learning.

InfoMax:

First postulated in information theory (Shannon, 1948), the maximum information principle presents a model of the transfer of information (Bell and Sejnowski, 1995). Infomax gained prominence in theoretical neuroscience as a testable framework for describing information flow from the external world to an internal representation of the world by maximizing mutual information between the two. Originally, internal representations were conceptualized as neuronal or cognitive, but any system that responds to external information must have some form of internal representation with properties that can be modeled by information theory (Rhee et al., 2012). Algorithms that build representations based on maximizing mutual information (MI) separate low MI, or independent, components of a set, and groups high MI, or most mutually interdependent ones. In the case of AMBA, the only available criteria to evaluate mutual information is gene coexpression. This is no guarantee of biological interactions since unrelated genes can be coexpressed, but coincidence is a sound precondition for interaction. InfoMax is central to independent component analysis (ICA) and a commonly-used objective in many other machine learning methods, such as variational autoencoders (Crescimanna and Graham, 2020; Hjelm et al., 2019).

Sparsity:

Living systems are complex, yet there is organization based on extreme resource constraints. Every molecular pathway of the cell essentially occupies the same space at the same time and each must be regulated while maintaining information flow. To achieve this, biological systems localize thermodynamic entropy, through membranes, maintained gradients and phase separations of reactive elements (Hyman et al., 2014). Dynamical properties of living systems arise from the links

between these compartments or molecular pathways (Furusawa and Kaneko, 2012). Too many links make information flow incoherent, while too few prevent feedback and regulation (Busiello et al., 2017). Sparse networks have evolved under these dual constraints to provide optimally controlled patterns of entropy within complex systems.

Representation learning algorithms have at least 3 areas to employ sparsity: the number of representations that are informative of the input data (feature or population sparsity), the samples of data that comprise any given representation (sample or lifetime sparsity), and the distribution of informative samples across all representations (dispersal) (Ngiam et al., 2011b). An ideal set of sparse representations would then be non-overlapping, with no representation carrying the majority of informativeness regarding the input data. This is in contrast to PCA, which derives components of descending variance, making the first few representations the most informative.

Sparse Filtering:

Sparse filtering (SFt) attempts to yield representations which are optimized with respect to both infomax and informativeness, the latter through the proxy of sparsity and the former by virtue of the transformations it imposes on its inputs (Zennaro and Chen, 2018). SFt uses a nonlinear transformation, l_2 normalization across features, l_2 normalization across samples and l_1 minimization. The only additional parameter for SFt beyond the input data is the number of representations to optimize over (Gultepe and Makrehchi, 2018). SFt is promising in that the transformations it applies preserve some aspects of the structure in the original data, namely cosine distance - points separated by a small cosine distance in the data space are mapped to points separated by a small euclidean distance in the feature space (Zennaro and Chen, 2018). Moreover the normalization across samples results in a minimal list of genes, a valuable feature when searching for biomarkers.

SFt has already shown potential for the analysis of biological signals and we sought to explore its utility in the high dimensional space of spatial transcriptomic datasets (Bruce et al., 2016). We generated representations from the Allen brain using PCA as the gold standard of the field, and using methods that feature Non-linearity (KPCA), InfoMax(ICA) and Sparsity (DLSC) as a basis of comparison for SFt. The survey of methods gave some unexpected results and suggests that for generating representations, no one method is best in all around. SFt derived representations however, were unique in their composition, presenting a variety of new applications and possibilities for this underappreciated approach to machine learning in spatial transcriptomics and other biological datasets.

RESULTS:

We established metrics for comparing representations derived from the AMBA dataset to ground truth anatomical information and to test the principles behind the algorithms that generated them.

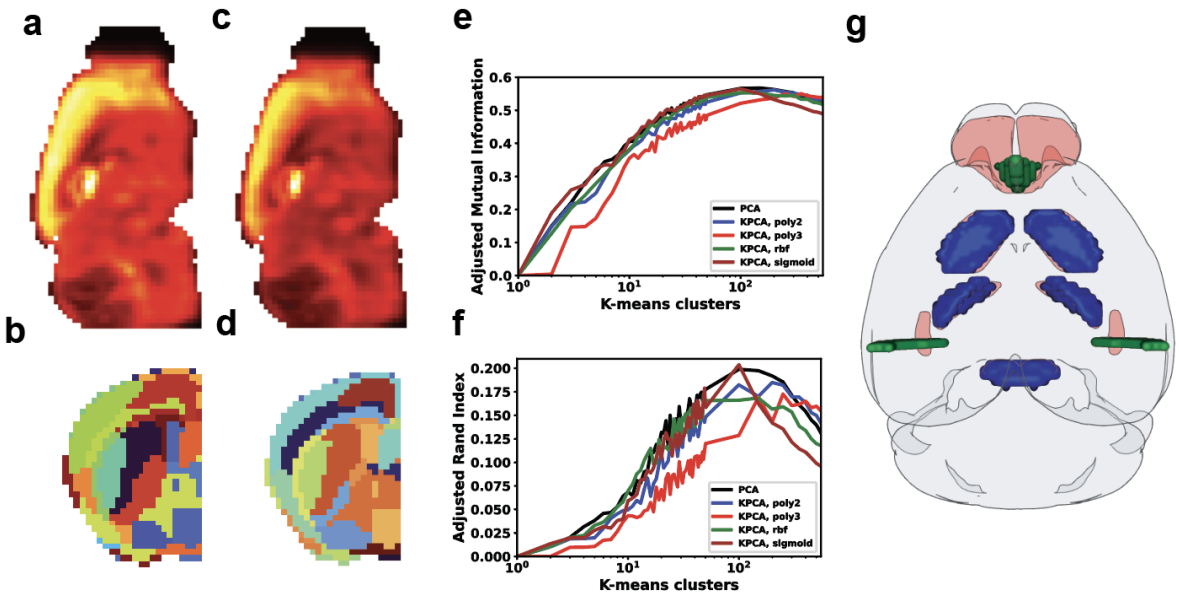

Figure 1: Linear and nonlinear PCA generate equivalent representations of anatomy from AMBA.
(a) First PCA component representation in a sagittal slice. **(b)** Representation of K-means clustering (K=200) of PCA components in a coronal slice of the left hemisphere. **(c)** First KPCA component with Quadratic kernel function in a sagittal slice. **(d)** Representation of K-means clustering (K=200) of KPCA components with Quadratic kernel function in a coronal slice of the left hemisphere.
(e,f) Adjusted Mutual Information and Adjusted Rand Index of clustered components using K-means clustering with K ranging from 1 to 50 and from 50 to 550 with a step of 50. The dimension reduction methods have been mentioned in the panel legends.
(g) 3D representation of K-means clusters of PCA components with selected top overlapping clusters (in blue) with brain regions (Olfactory Tubercle, Reticular nucleus of the Thalamus, and Pontine Gray) in red and low overlapping clusters (in green) with brain regions (Basolateral Amygdalar Nucleus, posterior part and Main Olfactory Bulb) in red.

Dimensionality reduction with PCA with k-means clustering generates representations of the AMBA dataset that are comparable to ground truth anatomical information.

In some of the earliest analysis of the AMBA data, SVD and k-means clustering was used to generate representations that resembled anatomy (Ng et al., 2009). We similarly applied PCA to generate 253 components (previously shown to represent 80% of the data variance) as a baseline for comparing more advanced methods (Ng et al., 2009). As shown previously, PCA of AMBA generated components that recapitulate

anatomy (**Fig 1a**), and k-means clustering on the full ensemble of PCA components further parcellated brain voxels into representations of anatomical ground truth (**Fig 1b**). We then asked if a nonlinear approach could more accurately reflect underlying biology and applied KPCA to the same dataset. Across multiple nonlinear transformations, the top ranked components remained visibly similar to those found by PCA (**Fig 1c**). Applying k-means to the components produced by KPCA produced similar parcellations, suggesting no advantage to the method despite KPCA's capacity to represent non-linearity (**Fig 1d**).

To quantify the performance of PCA vs KPCA, we calculated the mutual information between the output clusters of each method with segmented anatomy. We used mutual information (MI) and Rand index (RI) scores to test the accuracy of clustered representations against ground truth (**Fig 1e,f**). The Allen Brain Atlas labels 574 unique brain structures at 200 μm resolution. Even if every anatomical region could be defined by transcription, high numbers of clusters cause k-means to overfit the data, introducing artifacts and harming our performance metrics. To determine the optimal number of representations to fit anatomy, we generated MI scores from 1 to 550 k-means clusters. In all methods MI scores peaked well below 550, with PCA fitting optimally at 200 clusters. Nonlinear methods did not yield better performing representations at any number of clusters. The most accurate representations were predominantly subcortical, with only resolution preventing a more optimal fit to anatomy. Poor representations appeared as single slices or in areas that had artifacts from histology (**Fig 1g**).

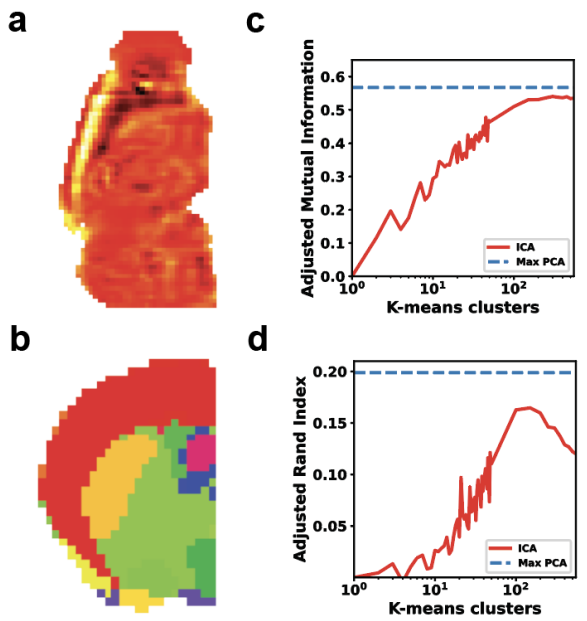

Figure 2:
Representations of ICA with 100 components;
(a) Selected active component in the cortex.
(b) K-means clustering of the components (K=200) in a coronal slice.
(c,d) Adjusted Mutual Information and Adjusted Rand Index of clustered components using K-means clustering with K ranging from 1 to 50 and from 50 to 550 with a step of 50.

ICA generates less accurate representations from the AMBA dataset.

With MI scores serving as a test of accuracy to anatomical ground truth, we asked if an unsupervised learning method that explicitly used MI across gene samples would yield greater accuracy based on the same metric. Using ICA alone, representations were comparable to PCA and KPCA when using the same parameters (**Fig 2a**). ICA + K-means gave parcellated representations that were visually similar to, though less detailed than, previous approaches when viewed from the coronal plane of the 3D dataset (**Fig 2b**). Many, however, were similar to low accuracy PCA and KPCA representations, fitting only within a single coronal slice (**Fig 2c**). While the AMBA is a 3D dataset, data was acquired in the coronal plane, raising the potential for slice-oriented artifacts from staining. Even when filtered for high spatial correlation between the two planes, ICA detected artifactual differences in staining across sections as information, and separated accordingly. MI and RI scores for ICA were the lowest of any method (**Fig 2d**).

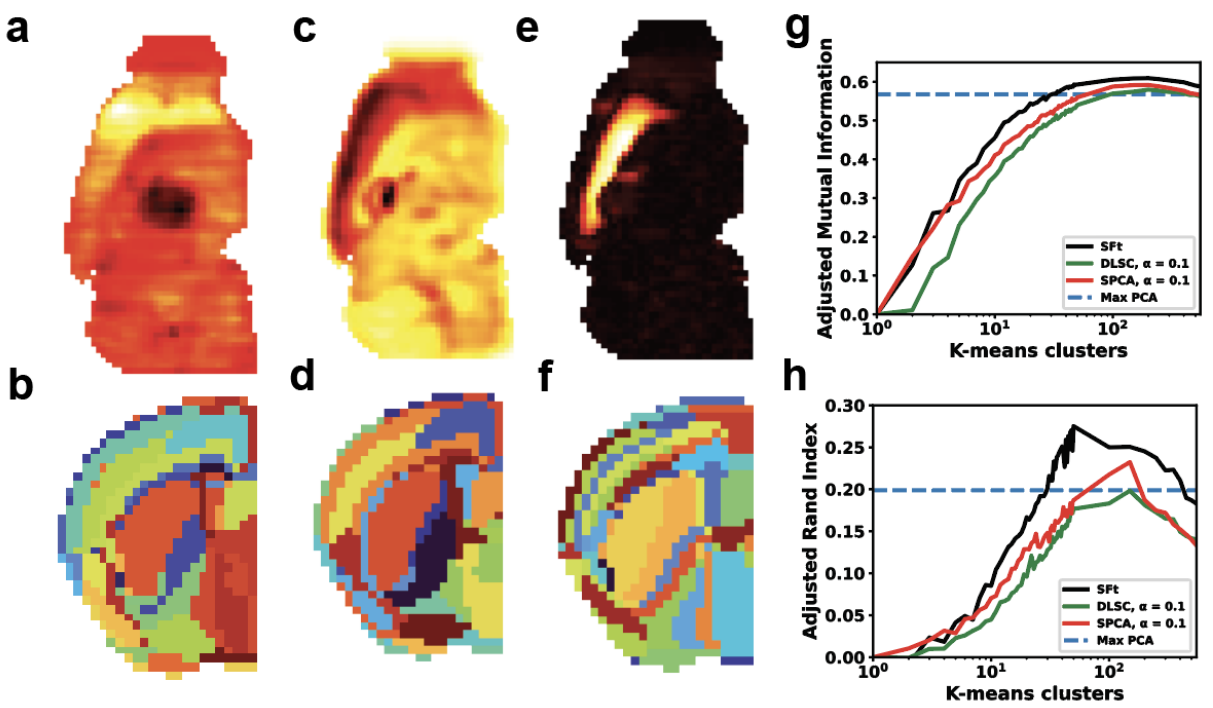

Figure 3: Sparse representations of AMBA.
(**a**) DLSC representation, active in the cortex. (**b**) K-means clustering (K=200) of DLSC representations in a coronal slice. (**c**) SPCA representation, chosen for similarity. (**d**) K-means clustering of SPCA representations. (**e**) SFt representation, active in the cortex . (**f**) K-means clustering of SFt components.

(**g,h**) Adjusted Mutual Information and Adjusted Rand Index of clustered components using K-means clustering with K ranging from 1 to 50 and from 50 to 550 with a step of 50.

Sparse methods offer finer representations, with peak results from SFt

Previous work has claimed better performing representations from AMBA using sparse methods (Li et al., 2017a). We generated representations starting with previously described optimal parameters, including a low sparsity value (α = 0.1). Unlike PCA, the ordering of sparse representations is not meaningful, but some covered similar anatomical features as other tested methods (**Fig 3a,c,e**). As previously reported, clustering of DLSC resulted in similar but visually finer representations than those generated by PCA (**Fig 3b**). Sparse PCA (SPCA) generated larger representations, but clustering gave comparably fine anatomical features to DLSC (**Fig 3b,c**)

SFt generated uniquely compact representations, called features, compared to other methods (**Fig 3e**). SFt + K-means clustering yielded fine features, including the clearest depiction of cortical layers of any method tested (**Fig 3f**). MI scores across all sparse methods showed greater accuracy for SFt than any other method, with a peak score of 0.61 at 200 clusters (**Fig 3g**). RI scores demonstrated an improvement over PCA with SPCA with SFt performing substantially better, peaking at 0.28 with 50 clusters. Visual inspection suggested improved fits were in part due to fewer coronal plane artifacts, but it could not be determined if SFt reached the limits of resolution.

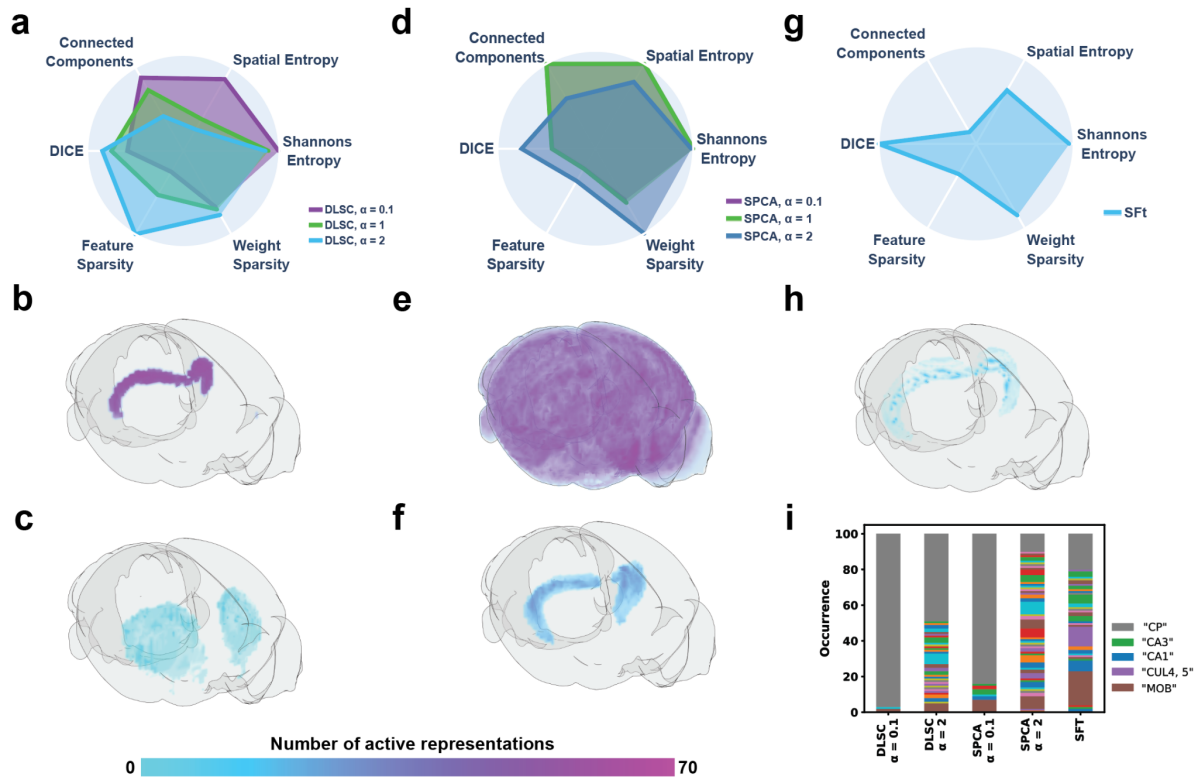

Figure 4: Relative performance metrics of DLSC, SPCA, and SFt. Representations derived from each sparse method without additional clustering were evaluated along 5 metrics with increasing sparsity values (α) when possible. (**a**) With increasing α, DLSC representations lose spatial entropy for increasing feature sparsity, with fewer connected components and improvements in individual fit to ground truth as measured by DICE score. (**b**) Map of the most accurately represented element of anatomy, by DICE score (CA3), at DLSC (α=0.1) shows a large number of overlapping representations. (**c**) At DLSC (α=2.0) less overlap occurs at the best represented anatomy (CP). (**d**) For SPCA with increasing α, weight sparsity increases while connected components shrink and DICE improves. (**e**) Anatomy best represented by SPCA (α=0.1) displays high overlap. (**f**) SPCA (α=2.0) has reduced overlap for top represented anatomy (CA3). (**g**) SFt has no α parameter, but balances lower spatial entropy and weight sparsity, with the fewest connected components and highest DICE scores of any method. (**h**) Anatomy most accurately represented by SFt (CA1) was still sparsely represented with the least overlap of all methods tested. (**i**) Labeling all representations by the anatomy they best represent showed some brain regions were redundantly represented by all methods, with CP overrepresented across all methods and α values.

SFt features are sparse representations of anatomical ground truth.

The uniquely sparse and compact features from SFt suggested anatomy with no further processing. We asked if SFt features were unique compared to other sparse methods and if they were accurate representations of ground truth. To better describe this qualitative difference, we developed metrics for 3D features and the gene

information used to build them: Spatial entropy or homogeneity of representation space, Shannon entropy or the total information, weight sparsity in the samples comprising each representation, feature sparsity or homogeneity within a given representation, DICE score or overlap between a representation and the best fitting anatomical ground truth, and the number of connected components or compactness of each representation. For DLSC and SPCA, we also tuned the sparsity parameter (α) to see how metrics would shift. Increasing α of DLSC predominantly increased feature sparsity while decreasing the connected components and overall spatial entropy, meaning information consolidated within a feature as it became more accurate (**Fig 4a)**. Looking at the top feature however showed that multiple representations overlapped with low α, which reduced with increasing sparsity (**Fig 4b,c**). In SPCA, increasing α also led to more compact representations and improved accuracy, however the increase in informativeness correlated with increasing weight sparsity, or the gene lists for each feature became more relevant to anatomy (**Fig 4d**). Low sparsity SPCA had the highest spatial entropy of any method with broadly diffuse and overlapping representations (**Fig 4e**). This improved with increasing α but SPCA retained the highest amount of feature overlap of all methods tested (**Fig 4f**). SFt balanced feature and weight sparsity with spatial entropy, but presented the most compact features tested with highest accuracy by DICE score (**Fig 4g**). SFt also had the least amount of overlap between features(**Fig 4h**). Despite a previously reported peak MI score at α=0.1, increasing α improved DICE scores, suggesting that our metrics were not reporting some aspect of representation learning. Noticing the high overlap of features at low sparsity, we asked which elements of ground truth anatomy were our representations fitting to. We were stunned to find that at low α, DLSC and SPCA were almost entirely deriving redundant representations of the caudoputamen (CP) (**Fig 4i**). Secondary clustering with k-means consolidated this overwhelming association, but the remaining information still managed to describe the rest of anatomy with better accuracy than PCA (**Fig 3g,h**). Increasing α better dispersed representations and allowed better individual fits to anatomy, while decreasing overall representation coherence (Li et al., 2017a). All methods and α values demonstrated similar amounts of Shannon's entropy, meaning a similar amount of information were conveyed, but no metric predicted spatially compact representations that fit optimally to ground truth by DICE (**Fig 4a-c**). SPCA at even higher α could offer comparable performance to SFt for specific representations, however SFt appears to provide the best overall distribution of informativeness through high DICE, MI and RAND index.

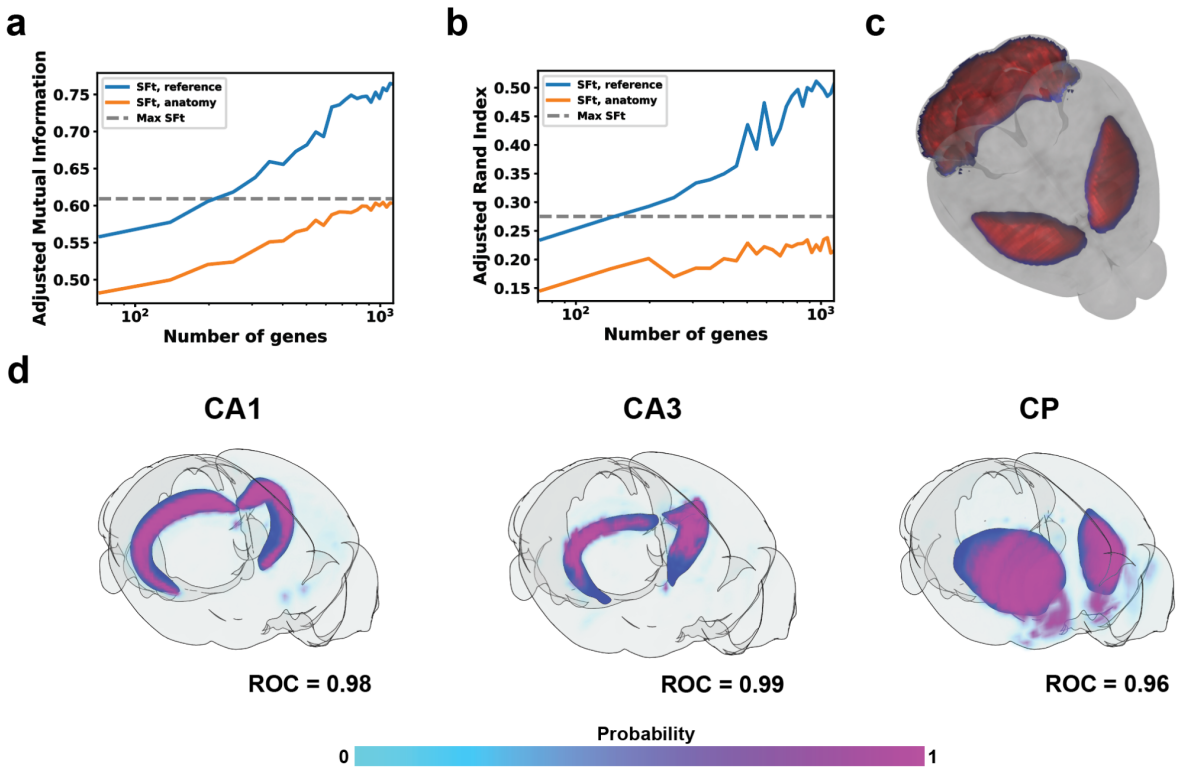

Figure 5: Performance metrics of compressed gene sets.
Lower ranked genes across all SFt representations were iteratively removed from AMBA data, followed by SFt/clustering on the compressed dataset. (**a**) Adjusted Mutual Information between the full dataset vs. compressed representations and anatomy vs compressed representations. (**b**) Adjusted Rand Index of compressed representations vs full representations and anatomy.
(**c**) Visualization of a selected feature derived from the SFt of the 2941-gene dataset (in red) with highest Pearson correlation to the 584-gene dataset (in blue), where 95% of the MI score with reference is met. (**d**) Ten genes with the largest associated weights were used to train classifiers for their corresponding brain region. Probability of positive classification is shown by the colorbar, and the corresponding anatomy shown in blue. ROC score for five-fold cross-validation is shown for each method.

Compressed SFt features can be used for robust classification.

Representations can render complex data into well classified labels, like cell types, pathological states, or in this case anatomy. The value of classification however is proportional to how labor intensive the starting data is to generate. The AMBA expression data took years to acquire and is unlikely to be repeated for any other strain of mouse, while experimentally available spatial transcriptomic methods can not offer the deep sampling of gene expression used here (Chen et al., 2015; Lewis et al., 2021). Given the already sparse gene compositions of anatomically relevant SFt features, we

asked if our input data could be compressed to the most relevant markers, while still representing anatomy.

We compressed the AMBA dataset by eliminating the lowest ranked genes from all representations and applied SFt to the remaining data, until only one top rated gene for a given representation was left. Based on MI and RAND, representation fidelity degraded very slowly for sparse methods (**Fig 5a,b**). MI and RAND index scores of compressed vs full datasets showed a near linear degradation, however compared to anatomical ground truth, representations from the degraded data did not see a performance drop until the gene samples fell were compressed to under 1000 genes and fell with a shallower slope, showing 95% fidelity to original ground truth measurements with only 584 genes, with inaccuracies limited to the borders (**Fig 5c**).

We then asked if the gene information from a single compressed SFt feature could be used to predictively label anatomy. SFt generates the most accurate representations with clustering, but for comparison we used the most commonly represented anatomy across all methods (**Fig 5d**). We used the top 10 genes from the best fitting representations for CA1, CA3, and CP to train a logistic regression based classifier. Data was randomized over 5 iterations, taking 80% for training and the remaining 20% and accuracy measured by receiver operating characteristic (ROC) curves. Classification was highly accurate, with scores above 95% for every classifier tested. This suggests that downstream applications with supervised learning can compensate for lower sample weight sparsity of a representation as long as it accurately captures ground truth.

DISCUSSION:

In a broad comparison of representation learning methods, most approaches could not outperform PCA with clustering for accuracy to anatomical ground truth, as measured by MI and RAND index between representation and ground truth. Most importantly, if a representation is merely being used to describe a known element of ground truth, a well-fitting PCA based representation can be just as actionable as any sparsity based method.

SFt was the only method tested that offered improved representations with consistently better performing metrics using secondary clustering. In addition, SFt features are themselves sparse, in that they are compact with minimal overlap, derived from a minimum number of samples and have well dispersed information across all representations. Perhaps most interestingly, SFt features alone are accurate representations of anatomical ground truth with no further clustering and once derived, only a fraction of the samples are needed to regenerate representations from data.

No method fully recapitulates biological ground truth.

While it was surprising that no method performed much better than PCA in fitting representations to anatomy, we should expect limitations to any unsupervised method. The adult mouse brain develops via spatio-temporarily defined genetic programs, using genes that have ceased activity long before histology occurs. There is no way to access that information through the AMBA and we are limited to genes that functionally define the adult brain at the time the data was collected. Brain plasticity also has a measurable role in morphology that cannot be reflected in the available data. Conversely, historical labeling of anatomical features was based on observed function, connectivity and morphology, with no consideration of gene expression. Finally as a practical consideration, many of the 574 labeled anatomical regions take up only a few voxels at the limited resolution afforded by the methods. Many of the fine regions of the mouse brain are simply lost or averaged into each other.

SFt representations are not explicitly optimized to be compact and contiguous, but reflect the properties of anatomical ground truth.

Sample lists from any representation learning method worked well for classification of anatomical ground truth using simple logistic regression, however the AMBA data is uniquely well annotated. In any other dataset, validating a representation presents a challenge. In PCA, lower ranked components are merely noise, while we identified many representations that depicted the same ground truth. All performance metrics suggest that SFt representations are inherently more accurate. SPCA was approaching equivalent accuracy by some metrics with increasing sparsity, but this may have to be determined empirically for any given dataset. While our basic implementation of SFt only offers the number of output representations as a controllable parameter, it can be tuned for further sparsity by reducing l_1 minimization to fractional values.

One obvious improvement that could be made to SFt for AMBA and other complex spatial dataset would be to consider neighboring voxels for unsupervised learning through spatial convolution. In this study spatial aspects of the dataset were implicitly retained and intrinsic to the representations are solely determined by gene expression. SFt representations appear more compact and contiguous than other methods only by virtue of their fit to ground truth. It is possible that even finer representations could be achieved by preceding SFt with spatial convolutional layers, however that may only be useful for datasets where spatial contiguousness is an important basis for learning.

Lossy compression for spatial transcriptomics.

SFt stands out from other representation learning methods for the minimal number of samples that comprise each feature. The representations generated here were even further compressed without significant loss of fidelity. In this pilot study,

after training the model on the 3k gene dataset we reduced the dataset to 580 genes and still generated a better fit to anatomical ground truth than other methods using a full dataset. The anatomy of the WT C57 B/6 mouse is well described, but much less so for knockout or transgenic mouse strains or for related species of interest. AMABA scale data acquisition is unlikely, but 580 genes are quite compatible with MERFISH or similar multiplexed spatial transcriptomic methods, meaning what once took years of work and millions of dollars could be largely recapitulated in a single session (Lewis et al., 2021).

Together these results raise the question of why SFt would represent ground truth better than other unsupervised methods when tested on AMBA. The method does not optimize for one single aspect of sparsity or mutual information and instead forces a balance between at times opposing aspects of data. This might not be an optimal strategy for every dataset, but it may reflect how biological systems themselves must balance infomax with sparsity to remain dynamic without losing coherence along molecular pathways. Future work should explore the sparse representations themselves as a means for finding biological mechanisms from large datasets.

METHODS:

Allen Mouse Brain Atlas (AMBA)

The Allen brain Atlas project (http://mouse.brain-map.org) has provided a comprehensive set of ~20,000 gene expression profiles with cellular resolution in the male, 56-day old C57BL/6J mouse brain (Lein et al., 2007). Image data were collected using the in situ hybridization method in sagittally-oriented slices with 200 μm inter-slice resolution and 25 μm thickness. The expression patterns were replicated in the coronal sections for ~4,000 genes of high neurobiological interest (Ng et al., 2009). The expression patterns were reconstructed in 3D and registered to a Nissl stain-based reference atlas (Allen Reference Atlas; ARA). The expression of a gene within 200 μm isotropic voxel is the average intensity of pixels in the pre-processed image called smoothed expression energy. Each 3D gene expression data consists of 67 × 41 × 58 (rostral-caudal, dorsal-ventral, left-right) spatially-matched volumes (Lein et al., 2007).

Data Pre-Processing

We followed the quality control measure implemented by (Bohland et al., 2010). The Pearson correlation coefficient was measured for each gene in coronal and sagittal experiments to find the higher-consistency dataset. Then 25% of genes with the lowest correlations were removed, resulting in a selection of 3,041 genes from which 2,941 still exist in the current version of the AMBA dataset, which we adopted for this study.

We extracted and concatenated the expression energy values in the left hemisphere for PCA and KPCA to form a large 31,512 voxel × 2,941 gene matrix, $E_{(v,g)}$, where v and g denote voxels and genes. We repeated the same procedure for the voxels in the brain (63113 voxels) for ICA, DLSC, Sparse PCA, and Sparse Filtering. We then z-transformed the data for each gene so that the mean of each gene's expression across the voxels was zero, and the standard deviation was one.

Principal Component Analysis (PCA)

As one of the most widely used matrix decomposition methods, PCA reduces the dimensionality of a dataset while preserving as much 'variability' (i.e., statistical information) as possible (Jolliffe and Cadima, 2016). We used the publicly available PCA package from the SciKit-Learn package in Python (Pedregosa et al., 2011). We chose 80% for the variance explained by the principal components, resulting in data transformation to a 253-dimensional space.

Kernel Principal component analysis (KPCA)

KPCA is the nonlinear extension of PCA, where nonlinearity is introduced by first mapping the data from the original input space into a higher dimensional feature space using a nonlinear map. Linear PCA is then performed on the mapped examples only in terms of dot products. Hence, the dot product space is determined implicitly by selecting a kernel function K that calculates the dot product between two input examples (Schölkopf et al., 1998).

The selected kernel functions for this study were the Quadratic, Cubic, RBF, and sigmoid. We implemented the KPCA object of the SciKit-Learn package (Pedregosa et al., 2011) for this study, with the criteria for dimension reduction in each analysis being the 80% explained variance.

Sparse PCA (SPCA)

The sparse extension of PCA looks for the sparse set of components that best reconstruct the data. The amount of sparseness is adjustable by the coefficient of the l_1 penalty, given by the parameter α. We used the SparsePCA package in SciKit-Learn (Pedregosa et al., 2011) for this study and chose $\alpha = 1$ for sparseness, and the number of components = 100.

Independent Component Analysis (ICA)

ICA is an unsupervised linear decomposition method that finds the underlying components and mixed sources in the data. It differs from PCA in that it uses the non-Gaussian structure of the data and has shown to be successful in analyzing multivariate data (Hyvärinen, 2013).

The FastICA algorithm of the open-source SciKit-Learn package (Pedregosa et al., 2011) was used for this project, choosing 100 components to be consistent with other methods.

Dictionary Learning and Sparse Coding (DLSC)

DLSC is a representation learning method that aims at finding a sparse representation of the input data (also known as sparse coding) in the form of a linear combination of basic elements. These elements are called atoms, composing a dictionary (Li et al., 2017b). The regularization parameter α controls the degree of sparsity of the sparse coding. In practice, the SciKit-Learn package for DLSC (Pedregosa et al., 2011) was used in this study with α values being {0.1, 1, 2}. The other hyperparameter, the number of dictionaries, was chosen 100 as is reported in (Li et al., 2017b) to yield the highest Mutual Information score with the mouse brain annotations.

Sparse Filtering (SFt)

Sparse filtering is an unsupervised feature learning method that efficiently scales to handle large input dimensions. The single hyperparameter to tune is the number of features. SFt optimizes for population sparsity (a few non-zero features represent each example), lifetime sparsity (each feature is active for a few samples), and high dispersal (features share similar statistics) (Zennaro and Chen, 2018).

The math proposed by (Ngiam et al., 2011a) considers a feature distribution matrix in a dataset where $f_j^{(i)}$ represents the jth feature value (rows) for the ith example (columns), resulting in $f_j^{(i)} = w_j^T x^{(i)}$. The first step is to normalize each feature by dividing it by its $l_2$norm across all examples $f\sim_j = f_j / \left\| f_j \right\|_2$. The features are then l_2 normalized across examples $f^{\wedge(i)} = f\sim^{(i)} / \left\| f\sim^{(i)} \right\|_2$. Then the l_1 penalty is used to optimize the features for sparsity so that in a dataset of M examples, the objective becomes: $minimize \sum_{i=1}^{M} \left\| f^{\wedge(i)} \right\|_1 = \sum_{i=1}^{M} \left\| \frac{f^{\sim(i)}}{\left\| f\sim^{(i)} \right\|_2} \right\|_1$.

We implemented the open-source GitHub repository of Sparse Filtering in Python (https://github.com/jmetzen/sparse-filtering) in this study. This software package transposes the voxel × gene matrix, $E_{(v,g)}$, to be consistent with the Matlab code provided by (Ngiam et al., 2011a). It also incorporated the soft-absolute activation function and L-BFGS minimizer. The values chosen for the number of features ranged from 100 to 650 with a step of 50.

K-means Clustering

After data representation using previously mentioned methods, we applied the unsupervised K-means clustering method. We used the publicly available K-means clustering implementation in SciKit-Learn (Pedregosa et al., 2011), and the number of clusters (K) ranged from 1 to 50, and 50 to 550 with a step of 50.

Similarity measures

We chose Adjusted Mutual Information (AMI) and Adjusted Rand Index (ARI) in the SciKit-Learn package (Pedregosa et al., 2011) to measure the similarity of clustered representations with anatomy, where the adjusted versions of MI and RI, despite the normal one, account for the chance.

We developed a function to compute the DICE similarity coefficient to measure the spatial overlap of brain regions and representations (see GitHub). It ranges from 0, meaning no overlap, to 1, showing perfect overlap. The formula to calculate DSC for two regions (A and B) is:

$$DSC_{(A,B)} = 2\,(A \cap B) \,/\, (A + B)$$

where ∩ is the intersection (Zou et al., 2004).

Logistic Regression

For Sparse Filtering, we found the candidate pairs of brain regions and features by measuring the DICE score of a given brain region against a thresholded feature. We thresholded the features using K-means clustering, with the single feature of interest as input, and choosing the number of clusters at two. We compared both labels against the region of interest using the DICE score, and kept the better of the two. We repeated this procedure for every brain region against all features, and kept the best region-feature combination.

We chose the brain regions CA1, CA3, and CP for logistic regressions, as they were among the top overlapping combinations for SFt and other methods. We determined the gene inputs for logistic regressions by identifying the largest ten weights by absolute value associated with a given feature produced by a transformation, and then finding the genes associated with these weights (**Table 1**). We then mean-centered and standard-scaled these ten genes, and prepared a shuffled five-fold cross-validation set. We trained a logistic regression from the normalized genes onto the region of interest, and reported the ROC score on the test set. Normalization, logistic regression, shuffling, cross-validation, and ROC score were all performed using their respective implementations in the scikit-learn library (Pedregosa et al., 2011).

Brain regions	CA1	CP	CA3
Genes	Fibcd1	Dock10	Pkp2
	Arl15	Serpina9	Tgfb2
	C630041L24Rik	Prox1	Tdo2
	Sstr4	Rasd2	Prox1
	TC1412430	Tdo2	Camk1d

	Matn2	Cyp7b1	Crlf1
	Cpne8	C78409	Il1rap
	Arhgap12	Rasgrp2	Lpl
	Dock10	Crlf1	Dock10
	Kcnd2	Pip5k1a	Sms

Table 1. 10 genes with highest weights for the overlapping SFt representation with the brain regions

Connected Components

We determined the number of connected components within a feature using the implementation of marching cubes in the scikit-image library (Pedregosa et al., 2011) to develop an adjacency matrix, and an algorithm that determines the number of connected components in a graph from its adjacency matrix available in the NetworkX library (Hagberg et al., 2008).

We reconstructed the representations generated by each method into the anatomical space, then applied the scikit-image implementation of marching cubes (Walt et al., 2014) to the resulting 3D array, using the mean value of the feature as the threshold level for the method. The output of this method is a set of vertices and faces that comprise a set of 3D surfaces. We constructed an adjacency matrix for the set of vertices using the set of faces, and found the number of connected components by passing this adjacency matrix to the number_connected_components algorithm in the NetworkX library (Hagberg et al., 2008).

Shannon Entropy

We measured the Shannon entropy of a feature by binning out the continuous values of the feature into discrete levels, and treated each discrete value as though it were a unique symbol in the Shannon entropy formula:

$$H(x) = -\sum_{i=1}^{n} P(x_i)\, ln\, P(x_i)$$

where each x_i in i=[1, n] is a discrete value of the binned feature.

Spatial Entropy

We developed a measurement of the spatial entropy of a feature using an extension of the grey level co-occurrence matrix method in use on 2D images. Pixel intensities present in an image are binned into discrete levels, and instances of co-occurrence of these levels is counted with specified spatial relationships as the elements of a matrix. We extended this method to a 3D image by including a three-dimensional relationship (Tsai et al., 2007) . Spatial entropy of an image can be

measured by finding the Shannon entropy of this matrix, in which each (i,j) index of the matrix is as though it were a unique symbol in the Shannon entropy formula described previously.

REFERENCES:

Bell, A.J., Sejnowski, T.J., 1995. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. Neural Computation 7, 1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129

Bohland, J.W., Bokil, H., Pathak, S.D., Lee, C.-K., Ng, L., Lau, C., Kuan, C., Hawrylycz, M., Mitra, P.P., 2010. Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy. Methods, Spatial gene imaging in the brain 50, 105–112. https://doi.org/10.1016/j.ymeth.2009.09.001

Bruce, N.D.B., Rahman, S., Carrier, D., 2016. Sparse coding in early visual representation: From specific properties to general principles. Neurocomputing 171, 1085–1098. https://doi.org/10.1016/j.neucom.2015.07.070

Busiello, D.M., Suweis, S., Hidalgo, J., Maritan, A., 2017. Explorability and the origin of network sparsity in living systems. Sci Rep 7, 12323. https://doi.org/10.1038/s41598-017-12521-1

Chari, T., Banerjee, J., Pachter, L., 2021. The Specious Art of Single-Cell Genomics. https://doi.org/10.1101/2021.08.25.457696

Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., Zhuang, X., 2015. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 348, aaa6090. https://doi.org/10.1126/science.aaa6090

Crescimanna, V., Graham, B., 2020. The Variational InfoMax AutoEncoder, in: 2020 International Joint Conference on Neural Networks (IJCNN). Presented at the 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207048

Furusawa, C., Kaneko, K., 2012. A Dynamical-Systems View of Stem Cell Biology. Science 338, 215–217. https://doi.org/10.1126/science.1224311

Gultepe, E., Makrehchi, M., 2018. Improving clustering performance using independent component analysis and unsupervised feature learning. Human-centric Computing and Information Sciences 8, 25. https://doi.org/10.1186/s13673-018-0148-3

Hagberg, A.A., Schult, D.A., Swart, P.J., 2008. Exploring Network Structure, Dynamics, and Function using NetworkX 5.

Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y., 2019. Learning deep representations by mutual information estimation and maximization. arXiv:1808.06670 [cs, stat].

Hyman, A.A., Weber, C.A., Jülicher, F., 2014. Liquid-liquid phase separation in biology. Annu Rev Cell Dev Biol 30, 39–58. https://doi.org/10.1146/annurev-cellbio-100913-013325

Hyvärinen, A., 2013. Independent component analysis: recent advances. Philos Trans A Math Phys Eng Sci 371, 20110534. https://doi.org/10.1098/rsta.2011.0534

Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374, 20150202. https://doi.org/10.1098/rsta.2015.0202

Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., Chen, Lin, Chen, Li, Chen, T.-M., Chi Chin, M., Chong, J., Crook, B.E., Czaplinska, A., Dang, C.N., Datta, S., Dee, N.R., Desaki, A.L., Desta, T., Diep, E., Dolbeare, T.A., Donelan, M.J., Dong, H.-W., Dougherty, J.G., Duncan, B.J., Ebbert, A.J., Eichele, G., Estin, L.K., Faber, C., Facer, B.A., Fields, R., Fischer, S.R., Fliss, T.P., Frensley, C., Gates, S.N., Glattfelder, K.J., Halverson, K.R., Hart, M.R., Hohmann, J.G., Howell, M.P., Jeung, D.P., Johnson, R.A., Karr, P.T., Kawal, R., Kidney, J.M., Knapik, R.H., Kuan, C.L., Lake, J.H., Laramee, A.R., Larsen, K.D., Lau,

C., Lemon, T.A., Liang, A.J., Liu, Y., Luong, L.T., Michaels, J., Morgan, J.J., Morgan, R.J., Mortrud, M.T., Mosqueda, N.F., Ng, L.L., Ng, R., Orta, G.J., Overly, C.C., Pak, T.H., Parry, S.E., Pathak, S.D., Pearson, O.C., Puchalski, R.B., Riley, Z.L., Rockett, H.R., Rowland, S.A., Royall, J.J., Ruiz, M.J., Sarno, N.R., Schaffnit, K., Shapovalova, N.V., Sivisay, T., Slaughterbeck, C.R., Smith, S.C., Smith, K.A., Smith, B.I., Sodt, A.J., Stewart, N.N., Stumpf, K.-R., Sunkin, S.M., Sutram, M., Tam, A., Teemer, C.D., Thaller, C., Thompson, C.L., Varnam, L.R., Visel, A., Whitlock, R.M., Wohnoutka, P.E., Wolkey, C.K., Wong, V.Y., Wood, M., Yaylaoglu, M.B., Young, R.C., Youngstrom, B.L., Feng Yuan, X., Zhang, B., Zwingman, T.A., Jones, A.R., 2007. Genome-wide atlas of gene expression in the adult mouse brain. Nature 445, 168–176. https://doi.org/10.1038/nature05453

Lewis, S.M., Asselin-Labat, M.-L., Nguyen, Q., Berthelet, J., Tan, X., Wimmer, V.C., Merino, D., Rogers, K.L., Naik, S.H., 2021. Spatial omics and multiplexed imaging to explore cancer biology. Nat Methods 18, 997–1012. https://doi.org/10.1038/s41592-021-01203-6

Li, Y., Chen, H., Jiang, X., Li, X., Lv, J., Li, M., Peng, H., Tsien, J.Z., Liu, T., 2017a. Transcriptome Architecture of Adult Mouse Brain Revealed by Sparse Coding of Genome-Wide In Situ Hybridization Images. Neuroinformatics 15, 285–295. https://doi.org/10.1007/s12021-017-9333-1

Li, Y., Chen, H., Jiang, X., Li, X., Lv, J., Li, M., Peng, H., Tsien, J.Z., Liu, T., 2017b. Transcriptome architecture of adult mouse brain revealed by sparse coding of genome-wide in situ hybridization images. Neuroinformatics 15, 285–295. https://doi.org/10.1007/s12021-017-9333-1

Ng, L., Bernard, A., Lau, C., Overly, C.C., Dong, H.-W., Kuan, C., Pathak, S., Sunkin, S.M., Dang, C., Bohland, J.W., Bokil, H., Mitra, P.P., Puelles, L., Hohmann, J., Anderson, D.J., Lein, E.S., Jones, A.R., Hawrylycz, M., 2009. An anatomic gene expression atlas of the adult mouse brain. Nat Neurosci 12, 356–362. https://doi.org/10.1038/nn.2281

Ngiam, J., Chen, Z., Bhaskar, S., Koh, P., Ng, A., 2011a. Sparse Filtering, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Ngiam, J., Koh, P.W., Chen, Z., Bhaskar, S., Ng, A.Y., 2011b. Sparse filtering, in: Proceedings of the 24th International Conference on Neural Information Processing Systems, NIPS'11. Curran Associates Inc., Red Hook, NY, USA, pp. 1125–1133.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Rhee, A., Cheong, R., Levchenko, A., 2012. The application of information theory to biochemical signaling systems. Phys Biol 9, 10.1088/1478-3975/9/4/045011. https://doi.org/10.1088/1478-3975/9/4/045011

Schölkopf, B., Smola, A., Müller, K.-R., 1998. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation 10, 1299–1319. https://doi.org/10.1162/089976698300017467

Shannon, C.E., 1948. A mathematical theory of communication. The Bell System Technical Journal 27, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Tsai, F., Chang, C.-K., Rau, J.-Y., Lin, T.-H., Liu, G.-R., 2007. 3D Computation of Gray Level Co-occurrence in Hyperspectral Image Cubes, in: Yuille, A.L., Zhu, S.-C., Cremers, D., Wang, Y. (Eds.), Energy Minimization Methods in Computer Vision and Pattern Recognition, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 429–440. https://doi.org/10.1007/978-3-540-74198-5_33

Walt, S. van der, Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T., 2014. scikit-image: image processing in Python. PeerJ 2, e453. https://doi.org/10.7717/peerj.453

Zennaro, F.M., Chen, K., 2018. Towards understanding sparse filtering: A theoretical perspective. Neural Netw 98, 154–177. https://doi.org/10.1016/j.neunet.2017.11.010

Zou, K.H., Warfield, S.K., Bharatha, A., Tempany, C.M.C., Kaus, M.R., Haker, S.J., Wells, W.M.,

Jolesz, F.A., Kikinis, R., 2004. Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index. Acad Radiol 11, 178–189. https://doi.org/10.1016/S1076-6332(03)00671-8