Parcellation of the mouse brain using molecular imaging

Alex Lee

Specific Aims

Parallel technological advances in molecular biology and microscopy have dramatically increased the ease with which experimental biologists can make spatially resolved measurements of large numbers of proteomic or transcriptomic features. These high-fidelity molecular mapping techniques are promising to catalog the enormous diversity of cell types in complex organs like the brain, whose cellular heterogeneity has been recognized since the early work of Ramon y Cajal [5]. These new capabilities are the focus of large scale collaborations facilitated by the NIH BRAIN initiative[17], which aim to precisely elucidate the structural organization of the brain by generating multimodal, cross-species datasets for comparison. The datasets that will result from these projects will be unprecedented in size and complexity. They will require new and innovative computational techniques to organize and analyze, but will transform our understanding of the molecular and cellular organizing principles of the brain.

Two principal focus areas in the neuroscience community's investigation of these relationships are variation in gene expression[21] and connectivity[18, 20], and how they interact in the circuitry of the brain. For example, single-cell RNA sequencing (scRNA-seq) has been enormously impactful for our ability to catalog neuronal variation[21], but to date it is unclear how graded transcriptomic variability observed in scRNA-seq overlays onto differences observed in higher order modalities such as function or connectivity. In Cembrowski et al. (2019)[6], for example, graded differences in connectivity in hippocampal subfield cells of the cornu ammonis (CA) correspond with connectivity and functional differences[7, 6]. Reconciling these perspectives would allow for a comprehensive account of the molecular components and configuration of the elements of neural circuits and comprise an important resource for experimental and systems neuroscience to the substrates of neural function.

One approach to understanding the composition of these circuits is to identify the major spatial components, or subregions detected in large scale datasets such as spatial transcriptomics. In this case a functional subregion can be defined as a distinct pattern of gene expression variation that can be used for categorization of a subgroup of cells. In the larger field of computational biology, a variety of unsupervised and supervised machine learning techniques have been developed for extraction of latent features for spatial clustering or subregion detection in spatial imaging data[27], but these methods typically target 2D tissue sections, and incorporate assumptions that are not ideal for analysis of 3D volumetric data or neural cell types, such as interactions only on a small spatial scale[24, 3, 35, 25]. In addition, there are few techniques suited for multimodal data integration of spatial imaging data, with most methods focusing on spatial transcriptomics exclusively.

I propose to address this unmet need by developing and applying new computational methods for machine learning-based spatial subregion detection and multimodal data integration to large scale datasets from the Allen Institute for Brain Science[31, 18, 14] and the Broad Institute[12]. I will use these techniques to map the structure and organization of these subregions in the mouse brain. I will also establish an interpretable statistical framework to map transcriptomic correlates and drivers of differential connectivity, leveraging the single-cell resolution of emerging spatial transcriptomic datasets. I hypothesize that continuous or graded gradients of gene expression that have been observed in well-characterized regions such as the hippocampus[6] are prevalent throughout the brain and can be more readily identified in spatial imaging data than in non-spatial datasets.

Aim 1: Establish a spatial transcriptomics data analysis pipeline to identify and characterize novel subregions in the mouse brain

First, I will develop a novel interpretable machine learning pipeline for unsupervised and interpretable discovery of functional subregions in 3D spatial transcriptomics data[12, 31]. I will validate this approach by testing whether this method can recover subfields of the hippocampus, specifically dorsal cornu ammonis regions CA1, CA2, and CA3, and the dentate gyrus. These subfields thus far have not been identified in a data-driven way in spatial imaging data. Next, I will characterize whether these subregions are driven by specific cell-types or by continuous variation within cell types using regression analysis. Determination of the principal gene expression spatial patterns in these datasets will pave the way towards understanding the organizing principles of cell type composition in the brain.

Aim 2: Determine transcriptomic and non-neuronal cell-type determinants of differential connectivity across the brain by integrating the Allen Mouse Connectivity Atlas with spatial transcriptomics datasets

First, I will a use a predictive regression model for association of connectivity patterns in the Allen Mouse Connectivity Atlas[18] with single-cell transcriptomes measured in Yao et al.[31] and Langlieb et al[12]. I will use this model to identify whether there are significant genes or gene modules that are predictive of differential connectivity among neurons. In addition I will identify whether this method can recapitulate an existing finding[10] that distinct transcriptomic classes of layer-5 pyramidal tract (PT) neurons are almost exclusively connected to one of either the thalamus or medulla. Systematic identification of the gene expression correlates of connectivity will give insight into the laws by which the brain constructs and constrains neural circuitry.

1 Background

From the first neuroanatomical studies[5], it has been recognized that although cells can be broadly grouped into high-level categories (like neurons, or glia), even visual inspection of silver-stained cells is sufficient to identify a diverse array of cellular subtypes. Originally, these cells were observed to primarily diverge based on their morphology and connection to other cells, however the development of new molecular techniques has revealed new axes by which these cells vary. Understanding the role of this diversity in facilitating the complex computations the brain performs remains one of the fundamental projects of modern neuroscience research. However, the scope and complexity of this project has increased in the modern 'omics age due to the accessibility of high dimensional molecular measurement tools. In particular, the ubiquity of high-fidelity molecular imaging data provides a unique opportunity to study the correspondence between our current and previous understanding of the organizing principles of brain structure and the patterns that can be identified in new datasets using statistical machine learning techniques.

Large scale molecular atlases such as the Allen Brain Atlas[14] (ABA) and Allen Mouse Connectivity Atlas[18] (ACA) have been critical tools to study the molecular organization of the brain. The ABA data uses one-at-a-time ISH to spatially map the expression of several thousand genes in the mouse brain.The ACA is a large scale imaging dataset where both cell-type specific and pan-neuron AAV vectors are used as tracers to map the connectivity of neurons to a source site, focusing on the right hemisphere. The integration of automated high-throughput data collection methods, comprehensive informatics, and fully open data sharing have facilitated an enormous number of studies (as measured by the 3,474 citations of the original Allen Brain Atlas Paper [as of 2023-02-18]). These datasets were integral to the development of the Allen Common Coordinate Framework (CCF)[28], a systematic effort to integrate gene expression, connectivity, transgenic expression, and histology to generate a neuroanatomical consensus structural labeling. This effort brought together a large team of neuroanatomists to develop a consensus spatial parcellation of the brain into known regions based on manual inspection of different data sources. The CCF is a highly important resource for neuroscientists, helping at the level of both hypothesis generation (for example, to identify regions of interest with respect to one or several molecular characteristics) or for hypothesis confirmation (to visualize distribution of certain features identified in an experiment). However, its creation was heavily manual, requiring a consensus of several expert neuroanatomists. Creating a new CCF from emerging datasets using novel technologies would require significant effort: therefore, development of computational techniques to organize and integrate new datasets in the mouse brain would be an important resource for neuroscientists.

New datasets utilizing new modalities are rapidly emerging in systems neuroscience, such as from Yao et al.[31], using MERFISH[8], and Langlieb et al.[12], using Slide-SeqV2[19], which offer spatial single-cell resolution. The opportunity to leverage single-cell level spatial measurements is transformative for the field, as it allows for the characterization of the organized spatial grouping of specific cell types into the components of neural circuits. The ability to analyze cells in their native spatial context is unprecedented, as previous spatially oriented investigations required techniques such as dissection of a particular region for bulk sequencing, dissociation for single-cell sequencing or cell-sorting using a particular cell marker. These techniques have relatively low spatial resolution and throughput. Alternative techniques with high spatial resolution such as in-situ hybridization[22, 14] precluded single-cell level measurement of many genes at a time. The development of techniques for high spatial fidelity measurements across the transcriptome[12] or with subcellular resolution[31] is particularly timely to resolve high degrees of observed heterogeneity within and across cell types discovered by non-spatial scRNA-seq[21]. For example, graded gene expression gradients within cell types have been observed in investigations of a small number of regions using techniques to target specific cell populations, particularly in pyramidal cells and in both excitatory and inhibitory neurons in the cortex[6]. Development of computational tools to incorporate this cellular heterogeneity and interpretably identify the major subtypes of cells in their spatial contexts will be crucial to understanding neural function[34].

Multimodal data integration is another key challenge for efforts to understand neural structural organization principles. A comprehensive classification of neuronal types is thought to require characterization involving both molecular (e.g. transcriptomic) and non-molecular modes such as connectivity[34, 6]. Systematic comparison of the cell type makeup and connection configuration of neural circuits would be a crucial step towards resolving function from structure in the brain. At the microscale, pioneering work such as from Chen et al.[9] has found that brain regions that have similar cell-type organization are likely to be connected. In a small number (~100) of single cells, an investigation by Economo et al.[10] has displayed a high correspondence between single-cell transcriptomic clusters and projection patterning in layer 5 pyramidal tract neurons. However, particularly with the release of one-hemisphere or whole-brain mouse spatial transcriptomic data, researchers can now systematically investigate the covariation of gene expression patterns and connectivity patterns. Several methods for multimodal data integration have been developed and applied for non-spatial multiomics datasets, such as MOFA[1], LIGER[29], although a ridge-regression workflow was shown to be effective[23] in modeling gene modules predictive of connectivity. These methods have yet to be tested for single-cell resolution datasets such as the above two, but could yield important insights as to the network organization of the brain.

Application of existing computational methods to integrate and identify major features in spatial transcriptomics datasets is limited by their large scale and sparse spatial structure. An attractive class of methods for spatial pattern discovery and multimodal data integration, for example for spatial transcriptomics data, utilizes Gaussian processes[25, 24]. These models are attractive because of their ability to incorporate prior knowledge on the form of the covariance between data observations in space. However, for 3D data with uneven spatial sampling it becomes harder to specify this relationship. For example, it is not obvious that the spatial correlation along the anterior-posterior axis should match, for example, the dorsal-ventral axis. In addition, Gaussian processes scale with O(n^3) and existing methods cannot easily scale beyond several tens of thousands of cells[24], making these methods unsuitable for region discovery in 3D whole-brain mouse transcriptomic datasets. An emerging perspective in modeling cells in space uses graph-neural networks[3]. In these models, a spatial cell neighborhood is created and networks can be optimized for supervised cell type prediction or using self-supervised learning (learning to predict characteristics of a hidden cell from its neighbors). However, these networks as implemented currently require construction of a spatial neighborhood of cells via thresholded distance. The construction of this cellular neighborhood graph in the brain is likely to incorporate many ad-hoc decisions; for example, it may not be desirable to have neurons that are spatially close be connected in a cell neighborhood graph, and it may be difficult to assign edge connections between neurons and various types of glia. A promising class of alternative methods are derived from non-negative matrix factorization[22, 30, 13] (NMF), which has previously been employed for spatial pattern identification in imaging data but not been adapted to the highly sparse, large scale setting of 3D spatial transcriptomics. A significant benefit of approaches derived from NMF is that it is possibly to easily visually interpret the factors.

In order to comprehensively catalog the transcriptomic and connectomic diversity of the mouse brain, the proposed research will create new computational methods that addresses these limitations. A comprehensive parcellation of transcriptome-defined subregions in the brain will be defined by developing a novel interpretable machine learning algorithm for latent variable discovery and then segmentation of the brain. By then leveraging the single-cell resolution (see Table 1 for a review of datasets used in this proposal) of emerging spatial transcriptomics datasets, I will analyze the specific gene gradients that contribute to these parcels. I will also utilize the single-cell measurements provided in these new datasets to conduct an association analysis of connectivity patterns and transcriptomic variation across the brain, to characterize whether gene expression gradients are attributable to differential connectivity patterns, as has been observed in some regions such as the hippocampus[2].

2 Aim 1: Establish a method for the parcellation and analysis of transcriptomic imaging data

My first aim will be to develop and test methods for unsupervised, interpretable discovery of functional subregions in the Langlieb et al.[12] and Yao et al.[31] whole-brain mouse spatial transcriptomics datasets. The size of these datasets (in Yao et al., at 10μm resolution, each gene is imaged in a large image of ~63 million voxels) requires development of robust software for subregion detection and characterization. Graded gene expression amongst cell types has been characterized in several regions in the brain, particularly in the hippocampus[6, 2], and provides an attractive target to validate the effectiveness of this approach. Completion of this aim will produce a clear enumeration of the principal gene expression gradients and their spatial location in the mouse brain.

2.1 Aim 1.1: Development and validation of an autoencoder approach for subregion discovery

I will build off of previous work performed by our lab[4] and others[30] that poses region discovery as a latent variable discovery problem. In both works, an NMF approach is used to identify spatial factors that are optimal for reconstruction. I am a co-author on the former work (Cahill et al.) where we applied this approach (which we call ontology-stability NMF, or osNMF, see Figure 1) to the Allen Brain Atlas dataset, however it has yet to be tested and refined on larger datasets such as the above spatial transcriptomics datasets.

An important modeling decision in factor analysis is the choice of inner dimension, or number of factors k. In order to guide this decision, we adopt the stability-driven framework of Wu et al.[30], where our algorithm is repeatedly fit on bootstrapped partitions of the data at different choices of k. The stability of the iterations is quantified using an Amari-like measure, which computes a distance between two factorization solutions through the sum of the absolute values of the row and column maxima of their cross-correlation matrix.

In a proof-of-concept analysis on the Yao et al. dataset I applied the osNMF technique. I conducted a sparse stability analysis, screening values of k between 5 and 25. This initial analysis identified an optimal value of 18 factors across the whole brain. Because of the large size of the dataset, I downsampled the dataset from a coronal-slice resolution of 10 μm to 40 μm. After binarizing the patterns to 18 for visualization, it was not clear by visual inspection that osNMF was able to resolve hippocampal subfields Figure 2.

Because of the inability of osNMF to resolve the desired hippocampal architecture, I introduced two changes. First, I implemented an autoencoder approach using 2-3 layers parameterized by 3D-convolutions instead of linear projection into low dimension, as in NMF. I include a ReLU activation in the last layer prior to a linear projection into native image spatial dimension for reconstruction. By constraining this last layer of neurons' weights to be strictly positive, and since the last layer activations are also constrained to be positive, we retain the interpretation of a "parts-based" positive representation that ordinary NMF is characterized. See Table 2 for a specific enumeration of the layers in the neural n etwork. When a network with this architecture was trained on downsampled image data (for comparison with osNMF), again with $k = 18$, hippocampal patterns were better resolved@fig-cnnpatterns. While the osNMF patterns do not resolve CA1 and CA3 (Figure 2), they are delineated in the autoencoder-derived patterns (CA3 appears as two patterns, which is supported by findings of strong gene expression gradients within pyramidal cells along CA3[6]).

With these initial results in mind, I will continue this analysis by applying the stability-driven autoencoder approach on both the Yao and Langlieb datasets at full spatial resolution. I will compare results from the different capture protocols, as Yao et al. was conducted with MERFISH and Langlieb with Slide-seqV2. I hypothesize that as Slide-seq provides coverage of the whole transcriptome, that we will be able to extract more accurate patterns from that datasets by comparing with CCF. I will compare the two sets of patterns with the CCF regions using the method in Cahill et al.[4] to quantify the correspondence of the discovered patterns with known brain regions by searching for best matches in the CCF ontology across scales and also compare with correspondence computed with patterns from osNMF.

2.2 Aim 1.2: Characterization of identified patterns using regression analysis and marker gene identification

In order to interpret the identified patterns, I will take several approaches. First, I will use a penalized regression model to analyze cell type contributions to each pattern. For a given pattern, I will regress the membership strengths (analogous to the weights in the factor or $\mathbf{W}$ matrix in NMF) on cell type counts at each location determined in either Langlieb et al. or Yao et al. For this I hypothesize Yao et al. will be more appropriate due to the subcellular resolution of MERFISH, whereas Langlieb et al. used deconvolution post-hoc to obtain single-cell profiles. By inspecting the coefficients attached to each cell type, I will be able to identify whether a given pattern is primarily driven by a single or a group of cell types. Based on findings from Chen et al.[9] using BAR-seq, I expect that the majority of patterns will be determined by gene expression heterogeneity within cell types that are indirectly driven by differences in connectivity across cells.

Second, I will compute last-layer activations (analogous to latent variable loadings in the $\mathbf{H}$ matrix in NMF) for each gene in the Langlieb (Slide-seq) dataset in order to obtain the top-10 most active genes for each pattern. From these I can cross-reference scRNA-seq profiles found by the Allen Brain Institute[21] to identify whether specific cell-types are marked by these spatially variable genes.

2.2.1 Concerns and alternate approaches

1. Autoencoder patterns are unstable: In initial testing, we observe visually a high degree of correlation between runs of the autoencoder until k becomes greater than about 25. However, preliminary stability analyses seem to indicate the settled value of k will be less than that. If it becomes necessary to fit a large inner dimension, we can reduce the degree of regularization or reduce the number of layers or parameters to induce greater stability empirically.

3 Aim 2: establish an analysis pipeline for the integration of transcriptomic and connectomic imaging data

Graded gene expression has been identified as a correlate of connectivity differences across a given subregion, with differences observed in the hippocampus, cortical neurons, and even non-neuronal cells[6]. In order to understand whether the relationships identified in small scale experiments with few cell types generalize, I propose a method based on penalized regression to identify gene expression correlates of differential connectivity in single cells across the brain, leveraging the high spatial resolution of the Yao et al. dataset.

I will first test the ability of my approach to identify previously determined relationships between single-cell transcriptomic variation in layer 5 pyramidal tract (PT) neurons and differential connectivity to the medulla or thalamus[10]. Economo et al. assayed connectivity using an AAV labeling procedure of layer 5 PT neurons specifically, whereas I will investigate whether connectivity and transcriptomic relationships can be identified reliably in two separate whole-brain datasets (Yao et al. and the Allen Mouse Connectivity Atlas developed by Oh et al.).

First, I will identify the *Slco2a1+* (medulla projecting) and *Hpgd+* (thalamus projecting) positive single cell clusters used in Economo et al. in the Yao et al. dataset (preliminarily, clusters “L5 ET CTX GLUT_1” [877 cells] and “L5 ET CTX GLUT_2” [766 cells]; note that ET refers to extra-telencephalic which is another term for pyramidal tract) using CARD[16], which annotates spatial transcriptomics-derived cell types using reference scRNA-seq atlases. The Allen Mouse Connectivity provides several measures of connectivity between two locations but I will use projection volume, which is defined as the sum of detected pixels of projection to a particular area normalized by both the number of pixels in that area and the total number of intense pixels for that injection. I will then regress projection volume or an indicator variable for connectedness in the medulla and thalamus (the two regions in Economo et al.) on MERFISH transcript counts for the two L5 PT neuron populations. This method is very similar to a method developed in Timonidis et al.[23] but has not been extended to single cell data. A differential expression analysis was conducted in Economo et al. for the two populations, identifying several genes that are in the MERFISH gene panel (*Npnt*, *Col8a1*, *Dmkn*, *Sorcs3*, *Cbln2*, *Cxcl12*, *Wipf3*, *Clmp*, *Syt6*, *Ctsc*) and so I will test the above model for its ability to identify these differences.

If succesful at delineating the differences in the two L5 PT populations, I will extend this approach to neurons across the brain. The division of neurons into classes and subclasses has been analyzed and gathered into a taxonomy by researchers at the Allen Brain Institute[32, 33], where the two PT classes in the previous analysis would be termed H3. H3 is the most granular cluster level, with H1 referring to a division of non-neurons, excitatory neurons, and inhibitory neurons and H2 referring to subclasses of those cells, such as L5 ET cells. For each subclass of cells I will regress projection volume against MERFISH transcriptome counts to determine on a per-cell type basis whether specific genes are correlated with differences in mesoscale connectivity, applying Bonferroni correction to threshold q-values at $q = 0.05$. I hypothesize that based on the results of Cembrowski et al.[6] that the majority of neuronal cell types will be associated with significant variation in connectivity. Once I have identified these genes I can perform gene-ontology analysis or pathway enrichment analysis to identify biological processes. I can also threshold the top-n of these by magnitude to provide likely marker genes for experimental validation by Retro-seq or ISH.

3.0.1 Concerns and alternate approaches

1. Low correspondence between projection targets in ACA and MERFISH cell segmentation locations: since the connectivity profiling in Oh et al. generated pan-neuronal (or specific Cre driver line-based) connectivity patterns in mice that are separate from the mouse measured in Yao et al., there is a concern that there is a low matching between cells in one dataset and the other, which would correspond to very little discriminative power to identify connectivity-gene relationships. In addition the AAV vector is not guaranteed to label all neurons and so there may be some stochasticity between of random chance in addition to image registration or overall cellular correspondence issues. My expectation is that this will be a relatively minimal effect–mouse-to-mouse variability was measured with repeat injections in Oh et al. for a different mouse, with the lowest $\mathrm{R}^2 = 0.85$ found in the nucleus of the optic tract. Overall the low amount of experimental noise in general suggests that if correspondence is too low at a given level of granularity, then I can average connectivity from a destination voxel to a group of voxels, for example those given by the very small "fine" subregions in the CCF.

2. Statistically correlations between connectivity and cell features in regression creates high risk of false positive gene associations: because of the statistical correlation between nearby cells, which presumably will violate the IID assumption of OLS, there is a high risk of false positives in the data. To remedy this I can create a spatial null distribution using Moran spectral randomization[26], an approach that is adopted in functional MRI studies to reduce the effect of spatial autocorrelation in association between spatial patterns.

3. Association between L5 PT neurons fails: if I am unable to resolve the validation pattern scheme I have set up, then I can switch to a scheme where I identify spatial patterns in the connectivity data and relate those to patterns I identfy in Aim 1 with correlation analysis or post-hoc spatial clustering, or discover them jointly in connectivity and transcriptomic data using an approach similar to Joint and Individual Variation Explained (JIVE)[15] or canonical correlation analysis[11].

4 Figures and tables

Table 1: Description of relevant datasets for the proposal and background information.

Dataset	Description
Allen Mouse Brain Atlas[14]	An in-situ hybridization (ISH) study of 20,000 genes across the brain. The dataset is generated "one-gene-at-a-time", meaning each gene is measured in a separate mouse. The resolution is fairly coarse, provided at 200μm isotropic resolution.
Allen Mouse Connectivity Atlas[18]	A dataset primarily using EGFP labeling across the brain to trace connectivity of different neurons and pan-neuronally. The EGFP is delivered via recombinant adno-associated virus (AAV) that was either implemented using a pan-neuronal expressed construct or using specific *Cre*-driver lines that are meant to express only in specific cell-types. The AAV construct operates by travelng from neuron to neuron in a anterograde manner, so that once injected, the researchers can wait several days and then sacrifice the mouse and image the brain to determine whether neurons in a target location are connected to the injection site. The resolution of this dataset is fairly high, at 25μm resolution in the transverse-imaging plane.
Langlieb Slide-SeqV2 dataset[12]	A Slide-seqV2 dataset generated across the brain. In Slide-seqV2 microparticle beads are spatially tagged and then applied to a slide which a tissue section can be applied to. Transcripts are captured through binding to a poly-T sequence applied the beads, which are also barcoded to allow for spatial array indexing. This technique is transcriptome wide, although capture efficiency is relatively low compared with traditional scRNA-seq. This dataset only covers one hemisphere of the adult mouse brain, although the resolution is quite high as the size of the beads is ∼10̌m.

Dataset	Description
Yao MERFISH dataset[31]	A MERFISH dataset generated across both hemispheres of the brain. In MERFISH, specific probe sequences along with a unique combinatorial barcode. During the imaging procedure, the barcodes are read off by sequencing-by-sequencing, where the barcode is designed so as to improve robustness in the decoding process. In Yao et al., 500 genes designed for discrimination between neuronal subtypes w as used to measure gene expression across both hemispheres of the mouse brain.

Table 2: Neural network architecture used in Aim 1.

Layer	Operation
1	Conv3D: 1x16 filters; 3x3x3 kernel; stride 2; dilation 2
2	BatchNorm + Dropout
3	Conv3d: 16x32 filters; 3x3x3 kernel; stride 1; dilation 2
4	BatchNorm + Dropout
5	Conv3d: 32x1 filters; 1x1x1 kernel; stride 1; dilation 1
6	Linear layer with neurons output dimension 18
7	Linear layer with neurons output dimension ~32M

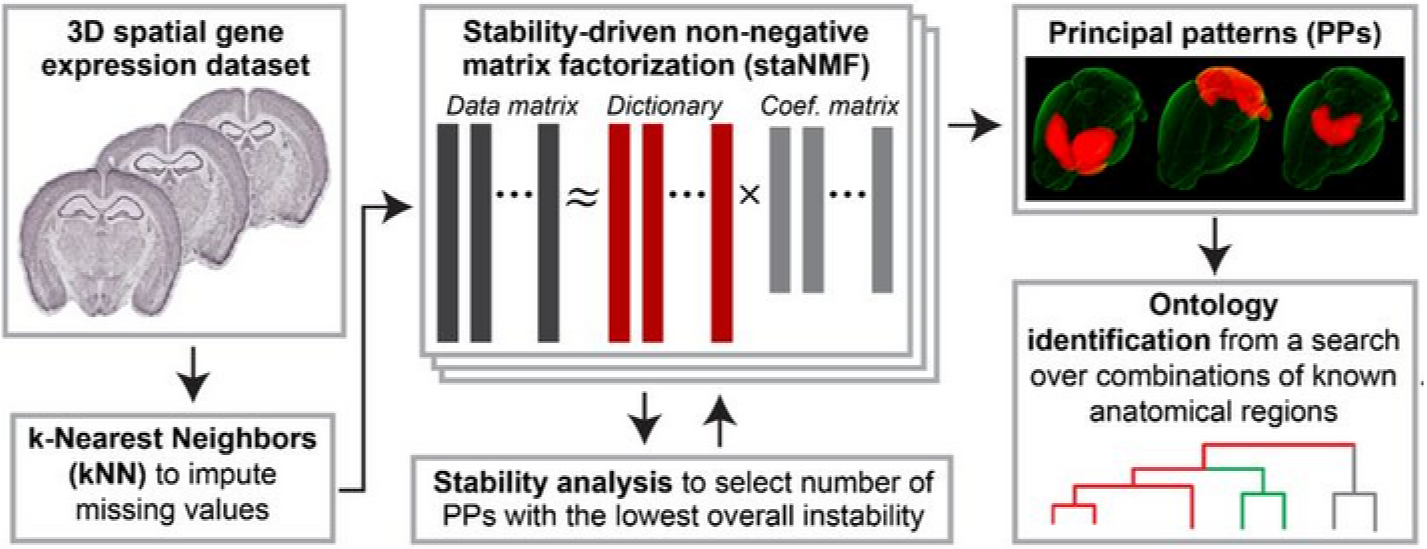
3D spatial gene expression dataset
k-Nearest Neighbors (kNN) to impute missing values
Stability-driven non-negative matrix factorization (staNMF)
Data matrix
Dictionary
Coef. matrix
≈
×
Stability analysis to select number of PPs with the lowest overall instability
Principal patterns (PPs)
Ontology identification from a search over combinations of known anatomical regions

Figure 1: Schematic of workflow for osNMF. A stability analysis is used to guide selection of the number of components k for NMF. Figure adapted from Cahill et al. (biorXiv, 2023)

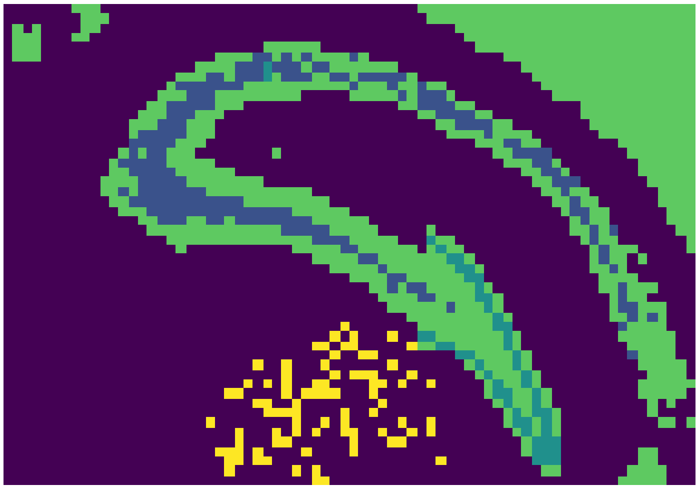

Figure 2: (a) NMF derived hippocampal factors

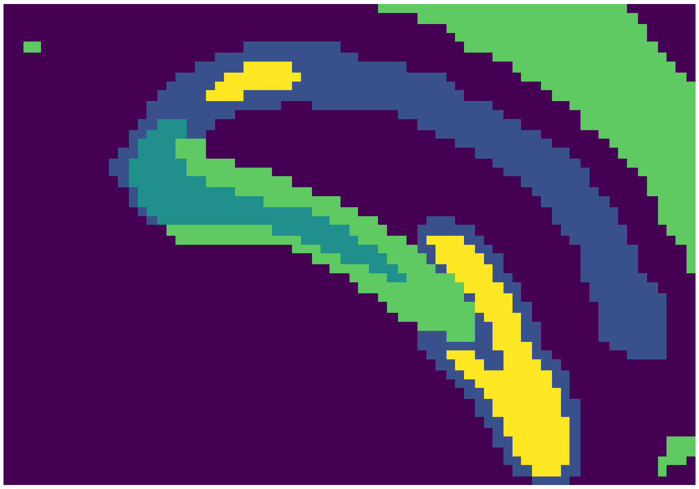

Figure 3: (b) CNN derived hippocampal factors

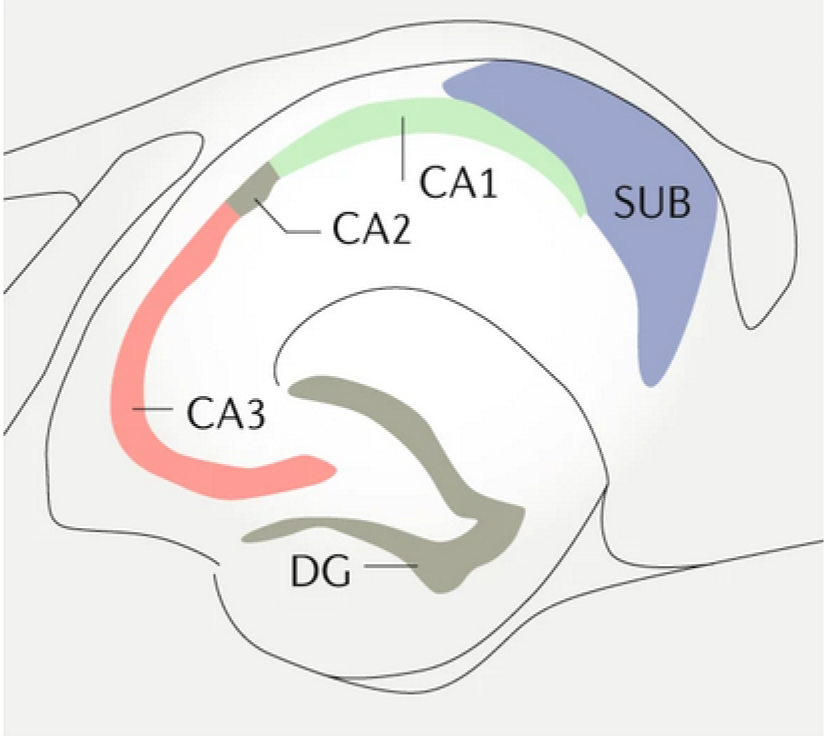

Figure 4: (c) Diagram of hippocampal subfield approximate anatomy, from Cembrowski et al. (2019) Results of preliminary test of osNMF and CNN-based autoencoder on the Yao et al. MERFISH data. A stability analysis was conducted and used to identify $k = 18$ factors using osNMF, and both CNN and osNMF models were fit at 18 factors. Patterns were binarized using the highest magnitude coefficient weight at each pixel. Shown in (a) and (b) is a zoom-in on the hippocampal formation with predictions for each method. (c) provides a reference from Cembrowski et al. (2019) of prospective hippocampal subfield anatomy.

References

[1] Ricard Argelaguet et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets". In: *Molecular Systems Biology* 14.6 (June 2018). Publisher: John Wiley & Sons, Ltd, e8124. ISSN: 1744-4292. DOI: 10.15252/msb.20178124. URL: https://www.embopress.org/doi/full/10.15252/msb.20178124 (visited on 04/05/2023).

[2] Michael S. Bienkowski et al. "Integration of gene expression and brain-wide connectivity reveals the multiscale organization of mouse hippocampal networks". In: *Nature Neuroscience* 21.11 (Nov. 2018). Number: 11 Publisher: Nature Publishing Group, pp. 1628–1643. ISSN: 1546-1726. DOI: 10.1038/s41593-018-0241-y. URL: https://www.nature.com/articles/s41593-018-0241-y (visited on 04/05/2023).

[3] Maria Brbić et al. "Annotation of spatially resolved single-cell data with STELLAR". In: *Nature Methods* 19.11 (Nov. 2022). Number: 11 Publisher: Nature Publishing Group, pp. 1411–1418. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01651-8. URL: https://www.nature.com/articles/s41592-022-01651-8 (visited on 02/15/2023).

[4] Robert Cahill et al. *Unsupervised pattern discovery in spatial gene expression atlas reveals mouse brain regions beyond established ontology.* Pages: 2023.03.10.531984 Section: New Results. Mar. 12, 2023. DOI: 10.1101/2023.03.10.531984. URL: https://www.biorxiv.org/content/10.1101/2023.03.10.531984v1 (visited on 04/05/2023).

[5] Santiago Ramón y Cajal. *Texture of the Nervous System of Man and the Vertebrates.* Vienna: Springer, 1999. ISBN: 978-3-7091-7323-7 978-3-7091-6435-8. DOI: 10.1007/978-3-7091-6435-8. URL: http://link.springer.com/10.1007/978-3-7091-6435-8 (visited on 02/15/2023).

[6] Mark S. Cembrowski and Nelson Spruston. "Heterogeneity within classical cell types is the rule: lessons from hippocampal pyramidal neurons". In: *Nature Reviews Neuroscience* 20.4 (Apr. 2019). Number: 4 Publisher: Nature Publishing Group, pp. 193–204. ISSN: 1471-0048. DOI: 10.1038/s41583-019-0125-5. URL: https://www.nature.com/articles/s41583-019-0125-5 (visited on 04/04/2023).

[7] Mark S. Cembrowski et al. "Dissociable Structural and Functional Hippocampal Outputs via Distinct Subiculum Cell Classes". In: *Cell* 173.5 (May 17, 2018), 1280–1292.e18. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.03.031. URL: https://www.sciencedirect.com/science/article/pii/S0092867418303118 (visited on 04/04/2023).

[8] Kok Hao Chen et al. "Spatially resolved, highly multiplexed RNA profiling in single cells". In: *Science* 348.6233 (Apr. 24, 2015). Publisher: American Association for the Advancement of Science, aaa6090. DOI: 10.1126/science.aaa6090. URL: https://www.science.org/doi/10.1126/science.aaa6090 (visited on 03/11/2023).

[9] Xiaoyin Chen et al. *Modular cell type organization of cortical areas revealed by in situ sequencing.* Pages: 2022.11.06.515380 Section: New Results. Nov. 6, 2022. DOI: 10.1101/2022.11.06.515380. URL: https://www.biorxiv.org/content/10.1101/2022.11.06.515380v1 (visited on 04/04/2023).

[10] Michael N. Economo et al. "Distinct descending motor cortex pathways and their roles in movement". In: *Nature* 563.7729 (Nov. 2018). Number: 7729 Publisher: Nature Publishing Group, pp. 79–84. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0642-9. URL: https://www.nature.com/articles/s41586-018-0642-9 (visited on 03/30/2023).

[11] Harold Hotelling. “Relations Between Two Sets of Variates”. In: *Biometrika* 28.3 (1936). Publisher: [Oxford University Press, Biometrika Trust], pp. 321–377. ISSN: 0006-3444. DOI: 10.2307/2333955. URL: https://www.jstor.org/stable/2333955 (visited on 04/07/2023).

[12] Jonah Langlieb et al. *The cell type composition of the adult mouse brain revealed by single cell and spatial genomics*. Pages: 2023.03.06.531307 Section: New Results. Mar. 8, 2023. DOI: 10.1101/2023.03.06.531307. URL: https://www.biorxiv.org/content/10.1101/2023.03.06.531307v1 (visited on 03/11/2023).

[13] Daniel D. Lee and H. Sebastian Seung. “Learning the parts of objects by non-negative matrix factorization”. In: *Nature* 401.6755 (Oct. 1999). Number: 6755 Publisher: Nature Publishing Group, pp. 788–791. ISSN: 1476-4687. DOI: 10.1038/44565. URL: https://www.nature.com/articles/44565 (visited on 04/05/2023).

[14] Ed S. Lein et al. “Genome-wide atlas of gene expression in the adult mouse brain”. In: *Nature* 445.7124 (Jan. 11, 2007), pp. 168–176. ISSN: 1476-4687. DOI: 10.1038/nature05453.

[15] Eric F. Lock et al. “JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES”. In: *The annals of applied statistics* 7.1 (Mar. 1, 2013), pp. 523–542. ISSN: 1932-6157. DOI: 10.1214/12-AOAS597. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3671601/ (visited on 04/07/2023).

[16] Ying Ma and Xiang Zhou. “Spatially Informed Cell Type Deconvolution for Spatial Transcriptomics”. In: *Nature biotechnology* (May 2, 2022), 10.1038/s41587–022–01273–7. ISSN: 1087-0156. DOI: 10.1038/s41587-022-01273-7. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9464662/ (visited on 04/06/2023).

[17] *NIH BRAIN Initiative Launches Projects to Develop Cell Atlases and Molecular Tools for Cell Access*. National Institute of Mental Health (NIMH). Sept. 22, 2022. URL: https://www.nimh.nih.gov/news/science-news/2022/nih-brain-initiative-launches-projects-to-develop-cell-atlases-and-molecular-tools-for-cell-access (visited on 02/15/2023).

[18] Seung Wook Oh et al. “A mesoscale connectome of the mouse brain”. In: *Nature* 508.7495 (Apr. 10, 2014), pp. 207–214. ISSN: 1476-4687. DOI: 10.1038/nature13186.

[19] Samuel G. Rodriques et al. “Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution”. In: *Science* 363.6434 (Mar. 29, 2019). Publisher: American Association for the Advancement of Science, pp. 1463–1467. DOI: 10.1126/science.aaw1219. URL: https://www.science.org/doi/10.1126/science.aaw1219 (visited on 03/11/2023).

[20] Yu-Chi Sun et al. “Integrating barcoded neuroanatomy with spatial transcriptional profiling enables identification of gene correlates of projections”. In: *Nature Neuroscience* 24.6 (June 2021), pp. 873–885. ISSN: 1546-1726. DOI: 10.1038/s41593-021-00842-4.

[21] Bosiljka Tasic et al. “Shared and distinct transcriptomic cell types across neocortical areas”. In: *Nature* 563.7729 (Nov. 2018). Number: 7729 Publisher: Nature Publishing Group, pp. 72–78. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0654-5. URL: https://www.nature.com/articles/s41586-018-0654-5 (visited on 04/05/2023).

[22] Carol L. Thompson et al. “Genomic Anatomy of the Hippocampus”. In: *Neuron* 60.6 (Dec. 26, 2008), pp. 1010–1021. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2008.12.008. URL: https://www.sciencedirect.com/science/article/pii/S0896627308010568 (visited on 04/04/2023).

[23] Nestor Timonidis, Rembrandt Bakker, and Paul Tiesinga. “Prediction of a Cell-Class-Specific Mouse Mesoconnectome Using Gene Expression Data”. In: *Neuroinformatics* 18.4 (2020), pp. 611–626. ISSN: 1539-2791. DOI: 10.1007/s12021-020-09471-x. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7498447/ (visited on 04/04/2023).

[24] F. William Townes and Barbara E. Engelhardt. “Nonnegative spatial factorization applied to spatial genomics”. In: *Nature Methods* 20.2 (Feb. 2023). Number: 2 Publisher: Nature Publishing Group, pp. 229–238. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01687-w. URL: https://www.nature.com/articles/s41592-022-01687-w (visited on 02/15/2023).

[25] Britta Velten et al. “Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO”. In: *Nature Methods* 19.2 (Feb. 2022). Number: 2 Publisher: Nature Publishing Group, pp. 179–186. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01343-9. URL: https://www.nature.com/articles/s41592-021-01343-9 (visited on 04/05/2023).

[26] Helene H. Wagner and Stéphane Dray. “Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods”. In: *Methods in Ecology and Evolution* 6.10 (2015). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12407, pp. 1169–1178. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12407. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12407 (visited on 04/07/2023).

[27] Benjamin L. Walker et al. “Deciphering tissue structure and function using spatial transcriptomics”. In: *Communications Biology* 5.1 (Mar. 10, 2022). Number: 1 Publisher: Nature Publishing Group, pp. 1–10. ISSN: 2399-3642. DOI: 10.1038/s42003-022-03175-5. URL: https://www.nature.com/articles/s42003-022-03175-5 (visited on 02/15/2023).

[28] Quanxin Wang et al. “The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas”. In: *Cell* 181.4 (May 14, 2020), 936–953.e20. ISSN: 0092-8674. DOI: 10.1016/j.cell.2020.04.007. URL: https://www.sciencedirect.com/science/article/pii/S0092867420304025 (visited on 02/15/2023).

[29] Joshua D. Welch et al. “Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity”. In: *Cell* 177.7 (June 13, 2019), 1873–1887.e17. ISSN: 1097-4172. DOI: 10.1016/j.cell.2019.05.006.

[30] Siqi Wu et al. “Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks”. In: *Proceedings of the National Academy of Sciences* 113.16 (Apr. 19, 2016). Publisher: Proceedings of the National Academy of Sciences, pp. 4290–4295. DOI: 10.1073/pnas.1521171113. URL: https://www.pnas.org/doi/10.1073/pnas.1521171113 (visited on 04/05/2023).

[31] Zizhen Yao et al. *A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain*. Pages: 2023.03.06.531121 Section: New Results. Mar. 6, 2023. DOI: 10.1101/2023.03.06.531121. URL: https://www.biorxiv.org/content/10.1101/2023.03.06.531121v1 (visited on 03/11/2023).

[32] Zizhen Yao et al. “A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation”. In: *Cell* 184.12 (June 10, 2021). Publisher: Elsevier, 3222–3241.e26. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2021.04.021. URL: https://www.cell.com/cell/abstract/S0092-8674(21)00501-8 (visited on 04/06/2023).

[33] Zizhen Yao et al. "A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex". In: *Nature* 598.7879 (Oct. 2021). Number: 7879 Publisher: Nature Publishing Group, pp. 103–110. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03500-8. URL: https://www.nature.com/articles/s41586-021-03500-8 (visited on 04/06/2023).

[34] Hongkui Zeng. "What is a cell type and how to define it?" In: *Cell* 185.15 (July 21, 2022), pp. 2739–2755. ISSN: 0092-8674. DOI: 10.1016/j.cell.2022.06.031. URL: https://www.sciencedirect.com/science/article/pii/S0092867422007838 (visited on 04/05/2023).

[35] Edward Zhao et al. "Spatial transcriptomics at subspot resolution with BayesSpace". In: *Nature Biotechnology* 39.11 (Nov. 2021). Number: 11 Publisher: Nature Publishing Group, pp. 1375–1384. ISSN: 1546-1696. DOI: 10.1038/s41587-021-00935-2. URL: https://www.nature.com/articles/s41587-021-00935-2 (visited on 04/04/2023).