

# Unsupervised Analysis of Gene Expression in Neurological Animal Models

Kevin Hu

*Massachusetts Academy of Math and Science at WPI*

February 8, 2017

## Table of Contents

<b>1 Abstract</b>	<b>5</b>
<b>2 Introduction</b>	<b>5</b>
<b>3 Literature review</b>	<b>7</b>
3.1 Neurological disease modeling . . . . .	7
3.2 Brain development . . . . .	8
3.3 Gene expression regulation . . . . .	9
3.4 Gene expression measurement . . . . .	10
3.5 Gene expression analysis . . . . .	12
3.6 The Allen Brain Atlas . . . . .	13
3.7 Neural networks . . . . .	14
3.8 Unsupervised learning . . . . .	20
<b>4 Research plan</b>	<b>21</b>
4.1 Researchable question . . . . .	21
4.2 Hypothesis . . . . .	21
4.3 Procedure . . . . .	21
<b>5 Methodology</b>	<b>22</b>
5.1 Overall workflow . . . . .	22
5.2 Software . . . . .	23
5.3 Data retrieval . . . . .	24
5.4 Data pre-processing . . . . .	24
5.5 Machine learning model . . . . .	25
5.6 Cluster identification . . . . .	29
5.7 Clustering comparison . . . . .	29
<b>6 Results</b>	<b>30</b>

6.1	Correlation clustering . . . . .	30
6.2	CAE-based encoding . . . . .	33
6.3	t-SNE . . . . .	34
6.4	Low-dimensional clustering . . . . .	35
6.5	Low-dimensional clustering comparison . . . . .	36
<b>7</b>	<b>Discussion</b>	<b>37</b>
7.1	Clustering validity . . . . .	37
7.2	ARI analysis . . . . .	42
<b>8</b>	<b>Conclusions</b>	<b>45</b>
<b>A</b>	<b>Limitations and Assumptions</b>	<b>49</b>

## Acknowledgments

The author would like to thank several individuals for their guidance and support during the process of this study. These include his advisor and STEM teacher, Ms. Siobhan Curran, who provided valuable guidance in the research process, and Mr. Jianwei Yang of Virginia Tech for his advice in the construction of the neural network. Several of the methods of validation employed in this study were suggested by or inspired by advice given by Dr. David Weaver and Dr. Oliver King of the University of Massachusetts Medical School. The review of this paper was further aided by Dr. Mahmoud Ghandi of the Broad Institute of MIT and Harvard through the MIT Program for Research in Mathematics, Engineering, and Science (PRIMES) program. The author would also like to acknowledge the encouragement and care provided by his parents, who provided the food, shelter, and electricity that made this research possible.

## 1 Abstract

Mouse models are fundamental to the current understanding of neurodegenerative disorders and serve as a readily available base for experimentation in drug development. However, less than 10% of drugs that have shown success in mouse-based preclinical trials show similar success in subsequent human clinical trials, which suggests that significant differences in gene expression are present at the molecular level. This study indirectly compared the developmental transcriptomes of mouse and human brains in order to further examine this inconsistency. Publicly available *in situ* hybridization and RNA-sequence data were compiled into species-specific datasets. Correlation networks were constructed in each dataset, and an optimized hierarchical method was used to identify clusters of inter-correlated expression. Cross-organism adjusted rand index analysis revealed a high degree of divergence, which contradicts the results of several previous analyses. Further examination using a deep manifold projection model revealed this divergence to be explained by widespread differences at the spatial and temporal levels. Immediate applications of this study include justification for the low success of mouse-derived treatments, a method of validation for future transgenic mouse models, and a robust process for deep analysis and visualization of high-dimensional transcriptomic data.

## 2 Introduction

The use of animal models is an essential component of the drug development process. Animal models, in particular mouse models, allow diseases to be studied and the effects of candidate drugs to be observed in real time. This method of elimination is crucial in neurological research, where mice serve as valuable filters to faulty drugs prior to clinical trials (Lin et al., 2014).

However, genetic differences between humans and animal models often confound drug tests. The protein products of a gene may interfere with the action of a drug, man-

ifesting as blockages of drug metabolism, mutated receptor proteins, and general physiological differences. Different patterns of spatial and temporal gene expression between humans and the model organism are factors in these interactions and are responsible for fundamental differences in morphology and cellular behavior (Burns et al., 2015). Previous studies have offered conflicting reports, such as stating that the brain is a highly conserved region Monaco et al. (2015) and Strand et al. (2007), while others have suggested that significant differences exist between the two organisms (Burns et al. (2015) and Lin et al. (2014)). These studies, however, have mainly focused on direct comparison of region-based comparison, which requires the assumption that these regions are truly homologous. A more definitive conclusion may be drawn if comparisons could be performed without the need for direct comparisons, and also with the consideration of expression over the entire developmental spectrum rather than only adulthood.

A possible solution for this process is to independently cluster two datasets of genes and to then compare them using a similarity statistic. Clustering, or the unsupervised categorization of data points into discrete sets, is widely used in gene expression analysis. In particular, clustering groups genes by similar patterns of expression, thus identifying shared biological pathways. Furthermore, clustering allows for the comparison of a specific gene's expression pattern in two different organisms by examining relative positions within each cluster. However, existing clustering algorithms rely heavily on measures of correlation that are prone to outlier leverage and high sample variance, both of which are common in biological samples (D'haeseleer, 2005). Gene expression data for both region and time are also high-dimensional, resulting in the data being excessively sparse. Furthermore, high variance in the data may prevent the algorithm from identifying discrete clusters. Recently, neural network algorithms have shown promise in the clustering of images such as those of the popular MNIST and USPS digit databases (Krizhevsky et al., 2012). Therefore, the goal of this project was to examine the differences between human and mice gene expression in cerebral development using a convolutional neural network that would facilitate the clustering of region-time expression matrices.

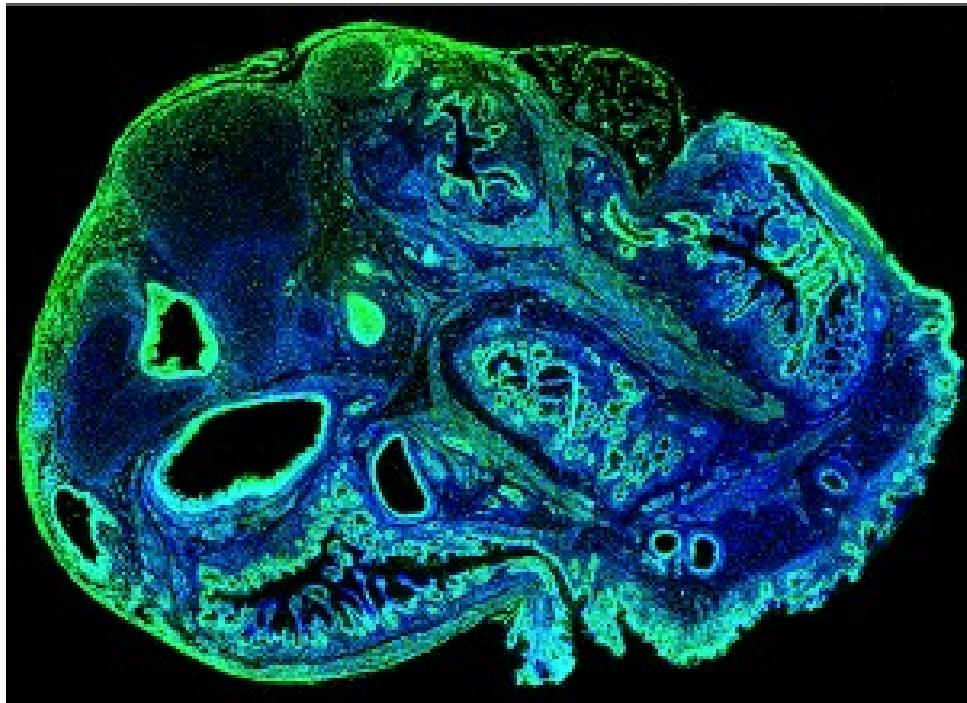
### 3 Literature review

#### 3.1 Neurological disease modeling

Mice are extensively used in neurological disease modeling due to their relatively short development cycle and the fragility and lack of available human brain tissue. Furthermore, mouse and human brains are highly conserved from an anatomical standpoint. However, recent studies offer conflicting views regarding the similarity of mice and human gene expression patterns, which casts doubt on the accuracy of results of mouse trials. Some studies suggest that neural development expression is highly conserved between mice and humans (Lin et al., 2014). Others suggest that fundamental differences between the two undermine the use of mice as a viable model for human diseases such as Alzheimer's and amyotrophic lateral sclerosis (ALS) (Burns et al., 2015).

With the advent of genetic engineering and regenerative biology, researchers have recently begun to switch to stem cells as a possible solution given the inaccuracies found in mouse models. Induced pluripotent stem cells (iPS cells) may be derived from nearly any patient cell sample and are therefore easy to obtain from noninvasive surgery. These iPS cells mimic embryonic stem cells (ES cells) and are able to differentiate into any adult cells. iPS cells may specifically be exposed to a variety of molecular and physical factors that guide them towards differentiation into neurons. These neurons, which are genetically identical to the donor, may then be used in drug discovery, disease modeling, and cell-based therapy techniques. In addition, these cells offer insights into the genetic backgrounds and effects of disease-associated genes by allowing precise *in vitro* examination of their underlying mechanisms (Imaizumi and Okano, 2014).

In addition to culturing neurons, researchers have also managed to guide development of entire clusters of neurons that may mimic the patterns and timings of human brain development (Lancaster and Knoblich, 2014). These cell clusters, or cerebral organoids, may be generated by culturing patient-derived iPS cells within an environment that encourages agglomeration and growth (Nguyen et al., 2015).

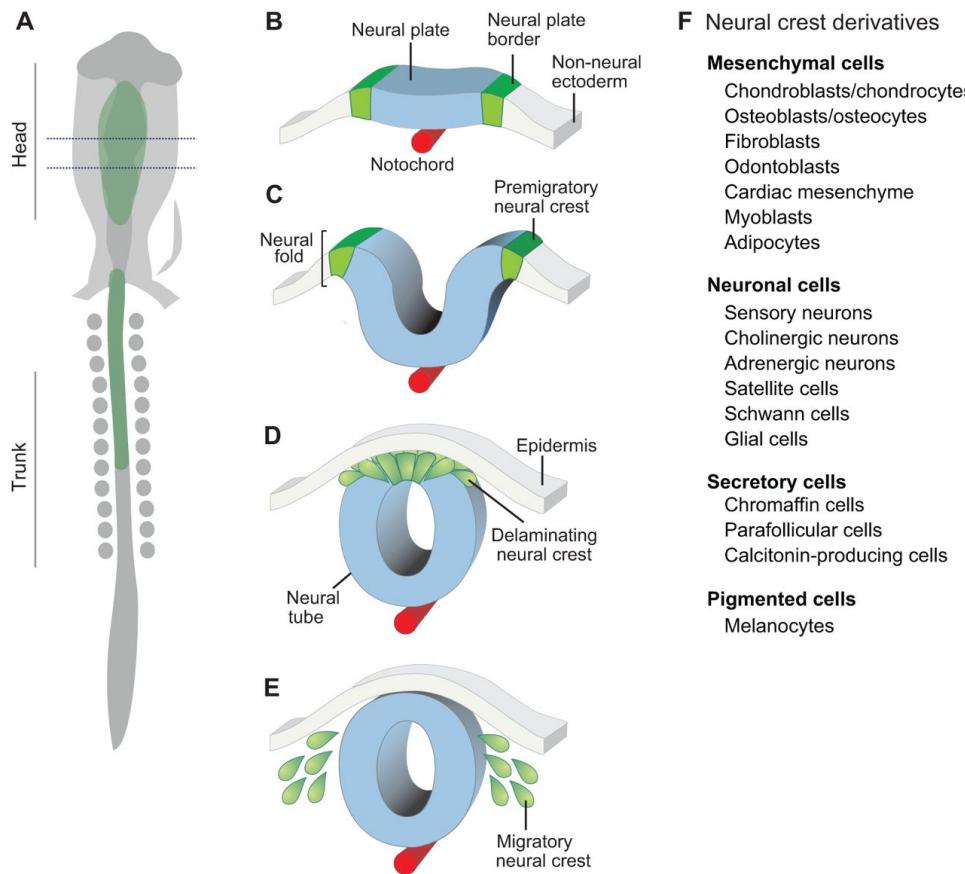


**Figure 1:** A cerebral organoid viewed under fluorescence microscopy (Nguyen et al., 2015).

### 3.2 Brain development

The mammalian brain is among the most complex organized structures in biology. The development of the animal nervous system, termed neural induction, is initiated by the formation of several primitive cell clusters early on in development. This initial differentiation of neural progenitors begins with the formation of the neural plate on the ectoderm (outer layer) of the embryo, which folds inward to form the neural fold/groove, creating the neural tube and crest (Figure 2).

The neural tube is characterized by four regions which develop into distinct adult structures (Bakken et al., 2015). The first of these, the prosencephalon, develops into the telencephalon and diencephalon, which further develop into the forebrain, hypothalamus, and eyes. The second region, the mesencephalon, develops into the midbrain. The third region, the rhombencephalon (named for its rhombus-like shape), is responsible for the lower regions of the brain such as the cerebellum, pons, and medulla oblongata, which originate from the metencephalon and myelencephalon. The last of these regions is the spinal chord.



**Figure 2:** Development of the neural tube (Simões-Costa and Bronner, 2015).

### 3.3 Gene expression regulation

Gene expression at the cellular level is regulated by a plethora of molecular means and is the foundation of an organism's development, growth, and survival. Although the vast majority of cells (except germ and certain immune cells) in an animal have identical DNA, their own structure and function is determined by these methods of regulation. This cellular regulation further determines the structure and function of the upper-level tissues, organs, and ultimately the organism itself.

Gene expression consists of the processes of DNA transcription to RNA and RNA translation to the amino acid chain that becomes the protein. Transcription is influenced by transcription factors, which affect the ability of an RNA polymerase (the enzyme primarily responsible for transcription) to bind to and form a transcription complex on a gene. Specific examples of molecules affecting gene expression include the addition of

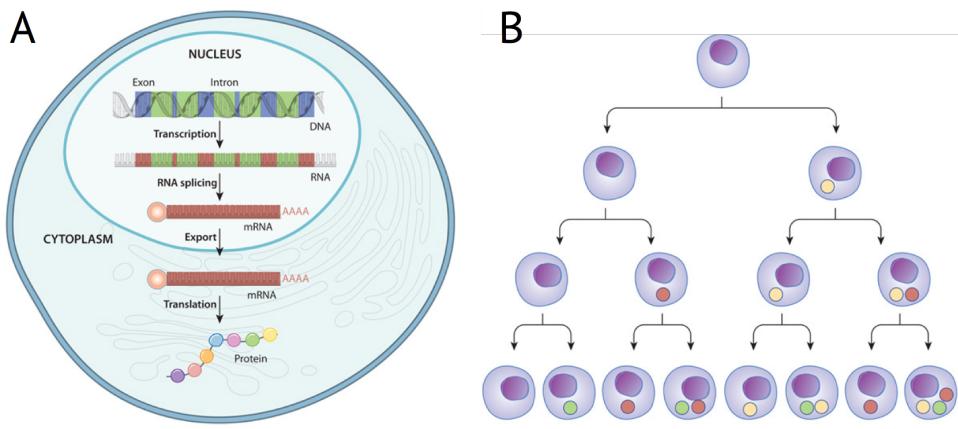
acetyl or methyl groups to the DNA and proteins that by binding to the DNA, can either attract or block the binding of RNA polymerase. In eukaryotic cells such as those of mice and humans, the RNA transcript must be further processed prior to transcription. This post-transcriptional processing involves the addition of cap and tail sequences that protect the sequence from digestive enzymes and allow for translation to occur. The RNA transcript also contains exon and intron regions; only the exon regions are incorporated into the final sequence to be translated. Once the RNA transcript has left the nucleus of the cell (through a nuclear pore), a ribosome (a large cellular machine that facilitates translation of RNA to protein) binds to the RNA and initiates construction of the amino acid chain. Once the amino acid chain is complete, the ribosome splits in two, and the protein may be further modified. For instance, the chain may be folded within a chaperonin, which oversees the formation of the correct conformational design for some proteins, or the protein may be tagged with small molecules that identify its destination. These processes, and many others, allow for extremely precise regulation of complex genetic pathways that are the drivers of embryonic development.

The embryonic phase of development is known for highly dynamic expression of genes across time and space, or spatiotemporal regulation. Differential gene expression over space and time is the feature that allows an organism to develop specialized cells. Conditional regulation of a select few high-level "master genes" can determine the fate of a cell. (Bisceglia, 2010)

### 3.4 Gene expression measurement

Numerous methods exist for the quantification and visualization of gene expression data. Two established methods include *in situ* hybridization (ISH) and RNA sequencing (RNA-seq).

ISH is a technique that allows specific sequences of nucleic acids to be located in a section of cells or tissue. This localization is accomplished through the principle

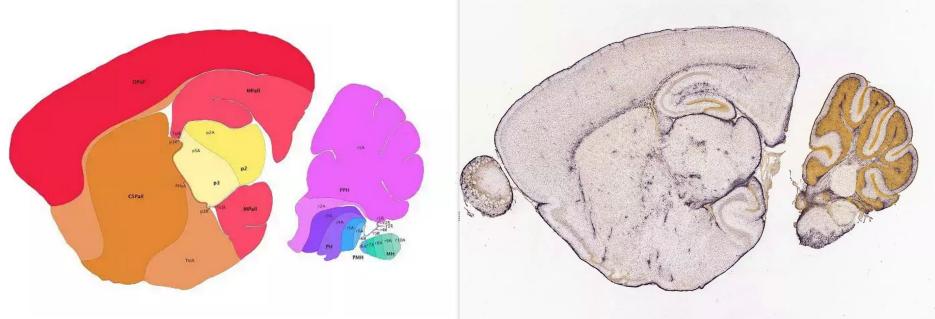


**Figure 3:** **A.** A general overview of the flow of information from DNA to RNA to protein. **B.** Demonstration of cell differentiation (Bisceglia, 2010).

that complementary nucleic acid sequences form hydrogen bonds (hybridize) with each other. The complementary sequence, or the *probe*, that detects the target sequence may be identified through fluorescent (fluorescent DNA components), radioactive (radioactive isotopes), or immunohistochemical (antibody-labeled) means. After the probe is allowed to hybridize to the target strand, RNases (RNA-digesting enzymes) hydrolyze any excess probes and the rest are washed off, leaving only bound probes behind. Fluorescent microscopy may then be used to highlight regions of gene expression, which may then be scanned and analyzed using image analysis algorithms (Angerer and Angerer, 1991). Regions of high expression of the sequence of interest are shown as darker areas in an image (Figure 4).

In addition to ISH, RNA-seq is also used in order to accurately measure gene expression. Because the primary function of RNA in the cell is as an intermediate in the transfer of information from DNA to protein, RNA expression levels may give an approximate indication of overall gene expression levels. In RNA-seq, as the name suggests, next-generation nucleic acid sequencing is applied to a collection of RNA sampled from a tissue region. This RNA sample is first isolated from the source tissue through digestion using detergents and enzymes. Once the RNAs are isolated, a next-generation sequencing machine reads the RNA to a digital storage device. These raw sequence data

are then formatted to identify exons and introns, which may be later used to identify gene expression levels. These gene expression values provide an approximate indication of the degree of a gene's expression within a cell or region (Wang et al., 2009).



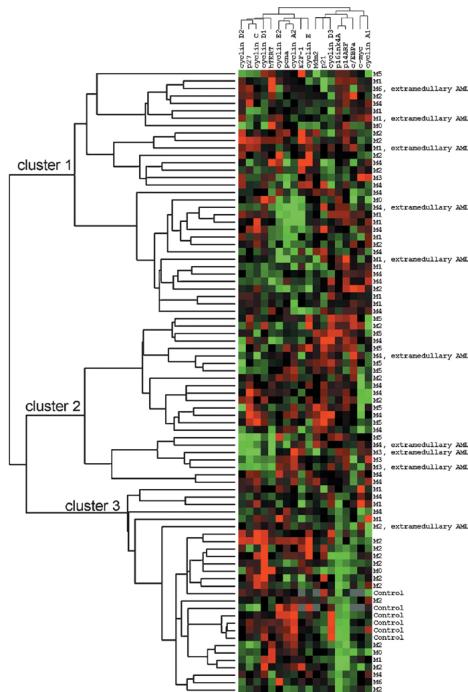
**Figure 4:** An ISH stain for glial fibrillary acidic protein (GFAP) RNA in the P14 mouse brain. At left is a diagram of the regions shown in the staining for reference (Allen Brain Atlas).

### 3.5 Gene expression analysis

Gene expression values offer valuable insights into the dynamics of regulatory networks and cellular processes. Analyses may be performed at any level of regulation described in Section 3.3. Among one of the most widely used expression analysis techniques are those which cluster genes by similarity of expression. The clusters obtained using these methods may be used to generate dendrograms (Figure 5) that provide a hierarchical ordering of the genes. Genes within a cluster are assumed to share a common underlying biological mechanism of action that is responsible for similar expression patterns.

The accuracy of a clustering depends heavily on the hyperparameter  $K$ , which denotes the number of clusters. In hierarchical clustering, a determining an optimal value of  $K$  is equivalent to determining the level at which the produced dendrogram is cut. An excessively low  $K$  may lead to large clusters which provide an overly general fit to the genes and fails to provide any meaningful insight. An excessively high  $K$ , on the other hand, may create more clusters than is necessary and creates categories where there are none. The validity of the clustering may be assessed by removing one of the predictor variables

(for instance, a specific timepoint or region of expression), and examining whether or not the resultant clustering is significantly altered (D'haeseleer, 2005). A significantly altered clustering would suggest that one variable affects the clustering to an abnormal degree, which may imply that the clustering method is chaotic.



**Figure 5:** An example of a clustering algorithm applied to expression values of genes related to acute myeloid leukemia. Dendograms were created for regions (top) and sample (left) (Muller-Tidow et al., 2004).

### 3.6 The Allen Brain Atlas

Founded in 2003 by Microsoft co-founder Paul Allen, the Allen Institute for Brain Science is a nonprofit research organization dedicated to the public pursuit of brain research. Among the many free datasets provided by the Institute are the Allen Developing Mouse Brain Atlas (located at [www.developingmouse.brain-map.org/](http://www.developingmouse.brain-map.org/)) and the Allen Atlas of the Developing Human Brain (located at [www.brainspan.org/](http://www.brainspan.org/)).

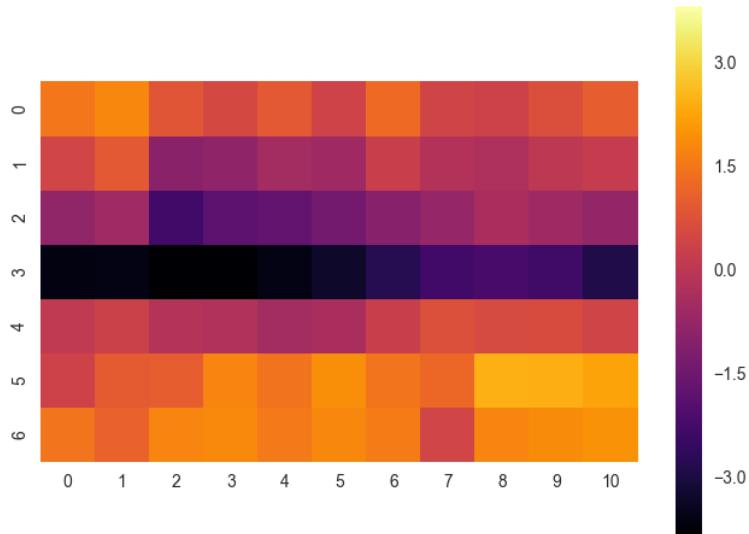
The Developing Mouse Brain Atlas features ISH data for over 2,100 genes across multiple stages of embryonic and postnatal mouse brain development. The raw images

of the ISH scans total 434,946 images and are created from ISH-stained slices produced using laser microdissection. These images were assembled into 3-dimensional voxel grids of expression values. An application programming interface (API) in the form of a bioinformatics pipeline allows researchers to access all data produced by the Developing Mouse Brain studies. The Developing Mouse API further allows researchers to obtain quantized expression values per region, which are calculated based off of analysis of the expression voxel grids. These expression values may be found for several different developmental time points, at several discrete brain regions, for each of the 2,100+ genes (Thompson et al., 2014).

Similar to the Developing Mouse Brain Atlas, the Developing Human Brain Atlas also features a diverse array of RNA-seq expression values for several thousand genes. These expression values are available for download as raw comma-separated-values (CSV) files from the Developing Human Brain Atlas Website, although an API is similarly available. In addition to RNA-sequencing values, extremely detailed and cellular-level magnetic resonance imaging (MRI) and microarray data are also included (Miller et al., 2014).

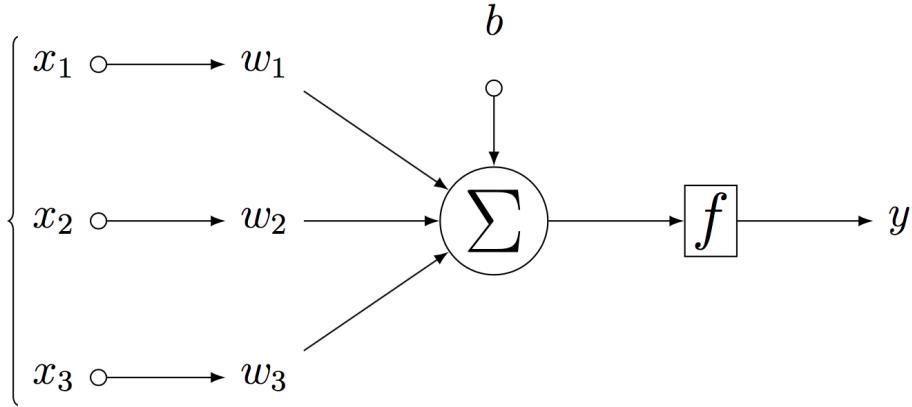
### 3.7 Neural networks

Neural networks are a subfield of machine learning, which is often cited as the “field of study that gives computers the ability to learn without being explicitly programmed,” a definition given by Arthur Samuel in 1959. Although invented around the mid 19-th century (Zilouchian, 2001), neural networks are one of the most flexible and powerful machine learning algorithms in modern computing. In several areas of research, neural networks offer state-of-the-art performance over traditional, hand-coded methods, such as in image pattern recognition, genomic analysis, and protein folding prediction. The basic neural network node, or *neuron*, is similar to its biological counterpart in that it receives a number of inputs and outputs an output (Figure 7). However, the two differ in



**Figure 6:** A heatmap of expression values for the Abelson murine leukemia viral oncogene homolog 1 (*ABL1*) gene in the developing mouse brain. Expression values were obtained from ISH stains and computed for each region using specialized 3-D voxel counts provided by the Allen Brain Atlas API. The y-axis is the development stage and the x-axis is the brain region (abbreviated). The expression values were transformed to a logarithmic scale to account for skew.

both structure and function. For instance, the biological neuron asynchronously receives a combination of excitatory and inhibitory impulses and sends a constant output signal. An artificial neuron instead is a mathematical construct that receives inputs synchronously and usually has a variable output value. In addition, whereas biological neural networks tend to be complicated in their topologies (many are not yet understood), artificial neural networks almost always have a tree-based structure of connectivity. The two further differ in their applications: whereas brains are capable of extrapolation and learning of novel tasks, artificial neural networks are specialized and cannot readily adapt to new tasks given learned information.

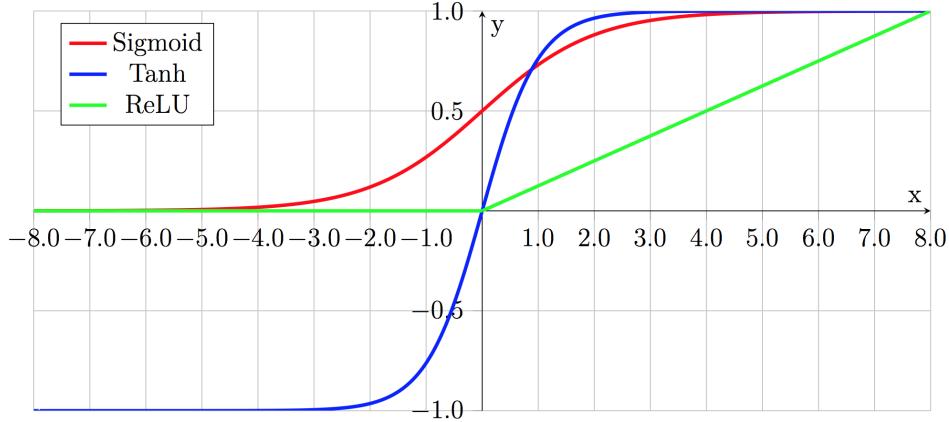


**Figure 7:** A diagram of a basic neuron with input values  $x_1, x_2$ , and  $x_3$  with respective weights  $w_1, w_2$ , and  $w_3$  and a bias  $b$ . The neuron computes the sum of these inputs and outputs a value  $y$  determined by activation function  $f$ .

Neural networks learn by minimizing an objective function which determines the degree of error present in the predictions of a network. This objective function is used to drive gradual minimization of the error value, which is accomplished by adjusting the specific weights and biases in a network. When a network is first initialized, the weights present are usually set to zero or randomly determined using a probability distribution function. As the network is fed data, the outputs are compared to target values and a loss value is computed. This loss value is then used to adjust the weights of the network to better fit the inputs to the output using a method called backpropagation. The effect of each neuron on the loss function is computed using the chain rule, which allows for the computation of the gradient of cost function. The weights are then updated using a method of gradient descent in order to minimize the loss value. This iterative updating allows the network to eventually learn a complex function given enough training data and time (Zilouchian, 2001).

There exist a multitude of activation functions which determine the output of an artificial neuron (Figure 8). Many of these functions are designed such that the loss gradients have smooth derivatives that can be handled by backpropagation. For instance, a step function would not be able to be backpropagated. The activation functions will

also partly determine the computational workload; a ReLU activation gradient, having one of two constant derivatives, requires significantly less processing power than a sigmoid activation gradient.

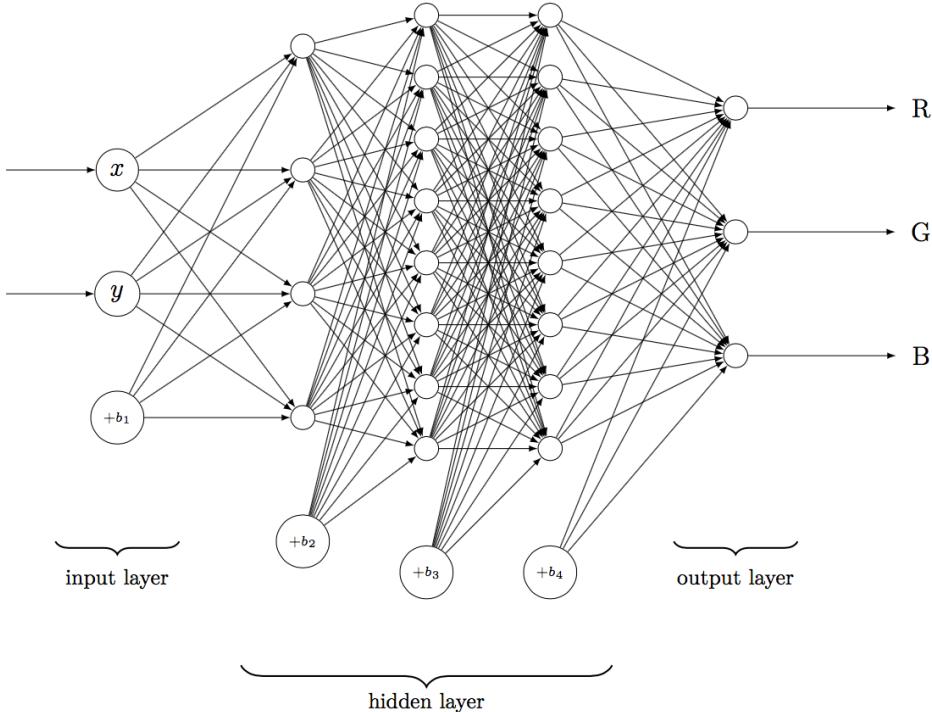


**Figure 8:** Graphs of common neuron activation functions.

**Table 1:** Equations of the activation functions shown in Figure 8.

Linear	Sigmoid	Tanh	ReLU
$a(x) = x$	$a(x) = \frac{1}{1+e^{-x}}$	$a(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	$a(x) = 0 \text{ if } x \leq 0, x \text{ if } x > 0$

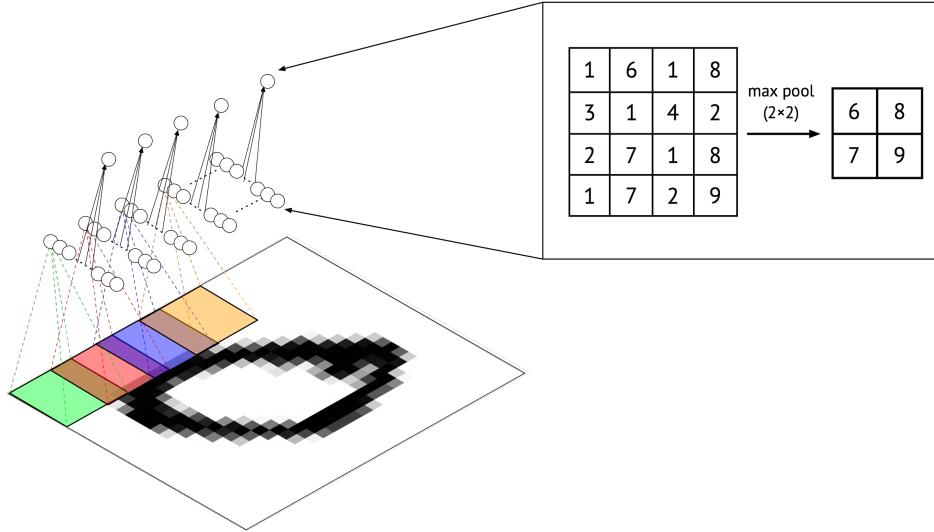
Neural networks are able to approximate complex functions such as classifiers through abstraction by multiple layers. In general, larger neural networks tend to yield more accurate predictions and lower losses. However, large networks are very computationally expensive to train and may also overfit the data when there are not enough training examples (Zilouchian, 2001).



**Figure 9:** An example neural network that learns to generate an image given  $x$  and  $y$  pixel position inputs. The inputs are the  $x$  and  $y$  coordinates of the pixel, and the outputs are the three red-green-blue (RGB) values that define the color of a pixel.

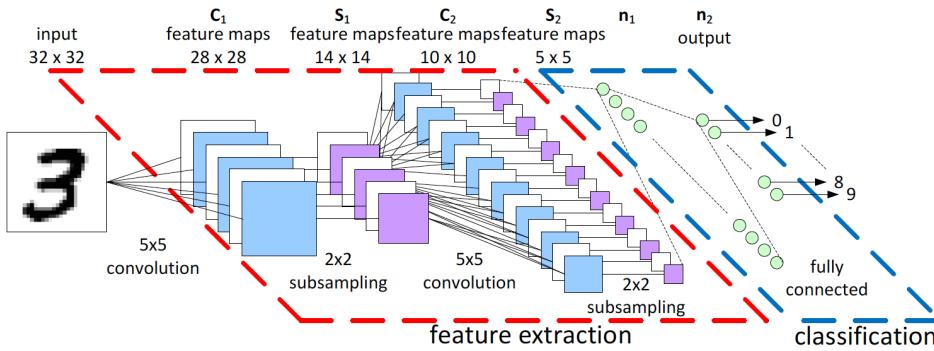
In image analysis, a specific type of neural network termed the convolutional neural network is often used. Convolutional neural networks are loosely based on the structure of the animal visual cortex, which contains overlapping receptive fields. Similarly, in the convolutional neural network architecture, the inputs to a neuron are shared, convolving the input image. In a convolutional layer, the inputs are processed using filters, which are the two-dimensional representations of the weights used by each unit, or kernel, of the layer. These filters are gradually adjusted during training to recognize specific features in the input. The results of the convolution are then processed using a pooling method (Figure 10) that produces a more abstract representation. In conventional neural networks, convolutional and pooling layers are often coupled in a sequence such that higher level features may be learned from the data. For example, while the first layer of convolution may be trained to recognize edges, later layers may recognize shapes and eventually discrete objects. The primary advantage of convolutional neural networks lie

in the fact that they are locally scale-invariant: convolution allows a feature of the image to be recognized regardless of position (Krizhevsky et al., 2012).



**Figure 10:** An example of a simple convolutional network designed for digit classification of the MNIST dataset.

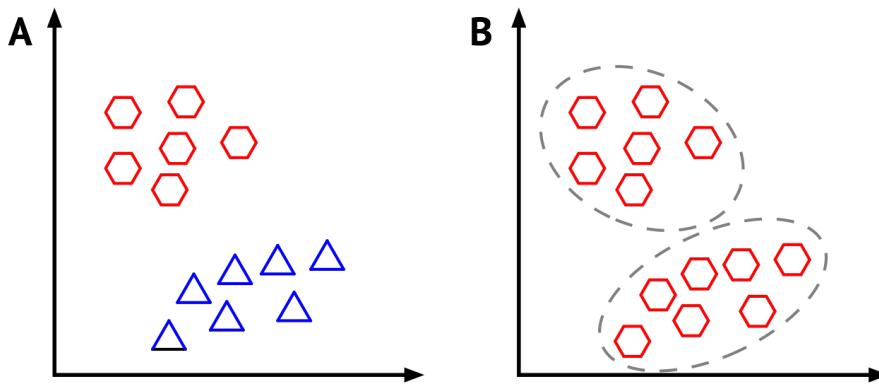
Following convolution and pooling layers, a dense layer of neurons is implemented to classify the features obtained (Figure 11). Variations of these networks are currently considered state-of-the-art in nearly all image classification tasks (Koushik, 2016).



**Figure 11:** An example of a multi-layer CNN for classifying handwritten digits from the MNIST dataset. The feature extraction section repeatedly convolves the input image to higher level features, which are fed as inputs to the classification stage. In this case, the classification stage is composed of two fully connected layers, with the final layer outputting the predicted digit classification (Peemen et al., 2011).

### 3.8 Unsupervised learning

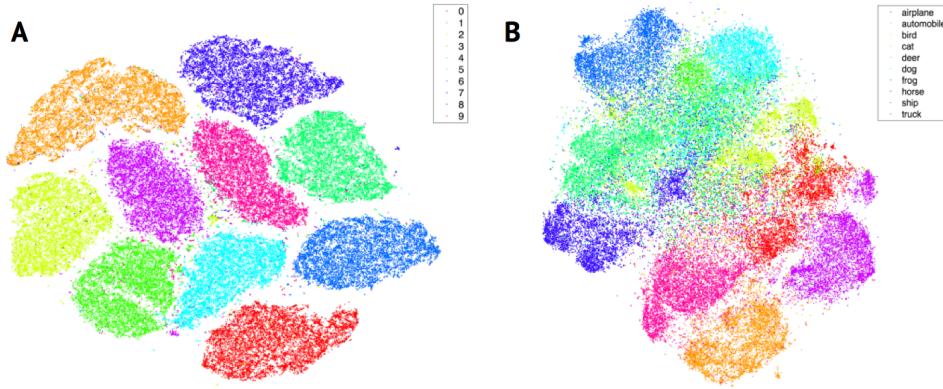
Machine learning may be divided into the two subfields of unsupervised and supervised learning. The primary difference between the two is in the training of the model: supervised learning trains the model using a predefined set of classifications, or *labels*, unsupervised learning trains the model without such guidance. Figure 12 gives a basic illustration of the difference between the two: in the left diagram, the datapoints are labeled as being red hexagons or blue triangles. With these labels, one could train supervised programs such as classifiers that predict the label of a data point given its x-y position. On the figure on the right, however, none of the datapoints are labeled (all of them are red hexagons). For these data, one could apply an unsupervised method such as a clustering analysis program to group datapoints by how similar they are to each other. This analysis allows researchers to make inferences regarding patterns and structures in the data.



**Figure 12:** An illustration of the primary difference between supervised (**A**) and unsupervised (**B**) learning.

Clustering is an essential method of unsupervised learning. Clustering analysis is often performed following dimensionality reduction methods that allow the data to be reduced to a simple two or three-dimensional representation while maintaining the original variance and structure in the data. Algorithms such as K-means, DBSCAN, and linkage clustering may then be employed to calculate clusters. In image classification

tasks, classification and dimensionality reduction allow the learned results of a neural network to be visualized (Figure 13). Furthermore, a convolutional neural network may be directly coupled with a clustering algorithm as a loss function in order to gradually produce high quality clusters as the network is trained (Yang et al., 2016).



**Figure 13:** The results of a *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE) dimensionality reduction on a clustering of the MNIST (A) and CIFAR-10 (B) datasets (van der Maaten, 2014).

## 4 Research plan

### 4.1 Researchable question

Does a significant divergence exist between the spatiotemporal gene expression patterns of developing mouse and human central nervous systems?

### 4.2 Hypothesis

Significant differences exist between the gene expression profiles of developing mouse and human central nervous systems.

### 4.3 Procedure

This project was entirely performed on a computer. Gene expression profiles for developing mouse and human brains were first downloaded from the Allen Brain Atlas.

Next, the data were formatted and transformed into sets of region versus time expression matrices into two datasets, a developing human dataset and a developing mouse dataset.

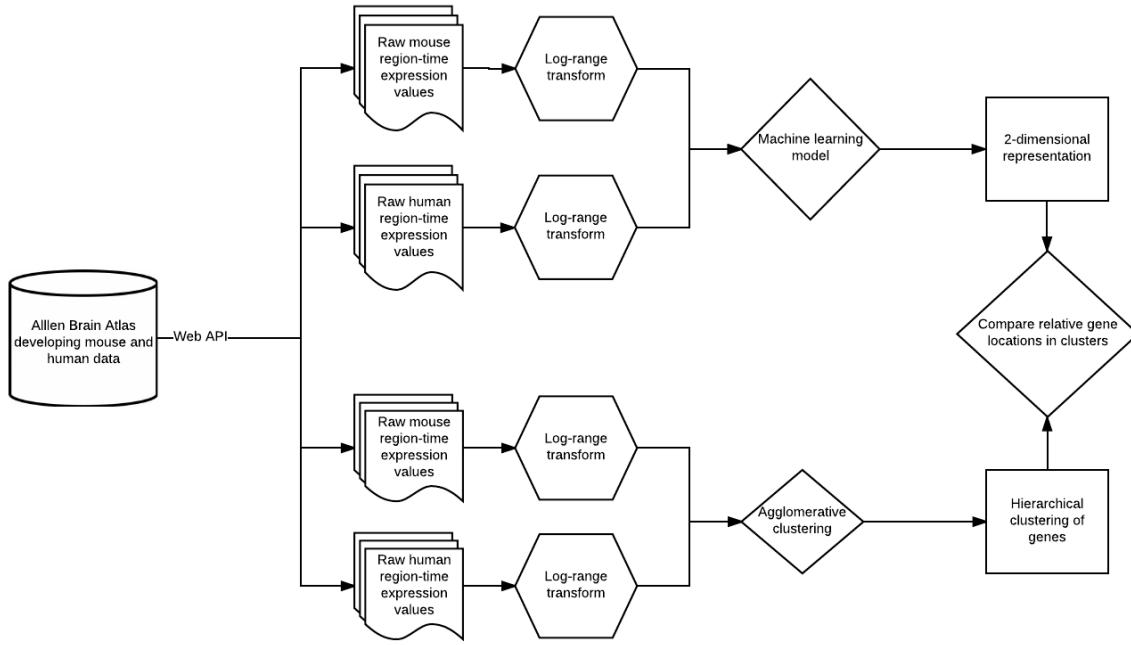
For each dataset, a convolutional neural network algorithm was applied in order to automatically classify genes into discrete clusters based on patterns of expression. The relative positions of genes in the human and mouse clusterings were examined. Different clusterings would suggest that the transcriptomic landscapes are different. The gene expression clusterings may be applied to highlight discrete drug-gene interactions in human versus mouse brains. Potential differing expressions may account for anatomical differences in mouse and human brains. As an extension, the similarity of expression patterns in cerebral organoids will also be considered. The insights from this project may also be applied to single-cell cancer sequencing, where clustering analysis may be used for the characterization and identification of stages of cancer cell development.

## 5 Methodology

### 5.1 Overall workflow

The overall process consisted of the retrieval, formatting, dimensionality reduction, and comparison of neural gene expression data. Gene expression profiles for both mouse and human donor brains were obtained from the Allen Brain Atlas website through either direct download (human) or a script (mouse). Expression values were log2-transformed and normalized to a (-1,1) range. A convolutional autoencoder (CAE) was constructed and embedded each 77 (mouse) or 150 (human) -dimensional expression matrix onto a four-dimensional manifold. t-distributed stochastic neighbor embedding (t-SNE) was then applied to further reduce the dimension count to two. This dimensionality reduction method was compared to traditional methods using only principal component analysis (PCA) and t-SNE in published literature. Density-based spatial clustering of applications with noise (DBSCAN) was then used to identify clusters in the two-dimensional space in

each mouse and human dataset. These clusterings were compared using the adjusted rand index (ARI) and mutual information score. In addition, this study also compared the CAE-based method with the results obtained by existing correlation-based hierarchical clustering of the raw expression values.



**Figure 14:** The overall process involved in this project.

## 5.2 Software

A number of open-source software packages and modules were used in this project. A script written in Ruby 2.0.0 was used to extract the mouse expression matrices from the Allen Brain Atlas API. Python 3.5.1 was used to perform all further analysis. Raw expression values for both the mouse and human datasets were manipulated using the Pandas package. Numpy was used for manipulation of the unlabeled expression matrices. The convolutional autoencoder was constructed with the aid of the Keras and Tensorflow (Abadi et al., 2015) frameworks. In addition, the NVIDIA CUDA 8.0 CuDNN v5 frameworks were used for GPU-based computation. The SciPy package was used for hierarchical clustering analysis, dendrogram construction, Pearson and Spearman correlation

analyses, and distance matrix construction. The scikit-learn package was used for t-SNE, DBSCAN, PCA, and computation of the adjusted rand indices and mutual information scores. Graphs and other visualizations were constructed using the Matplotlib, Seaborn, and Plotly libraries. Gephi and D3.js were used for visualization of the correlation graphs.

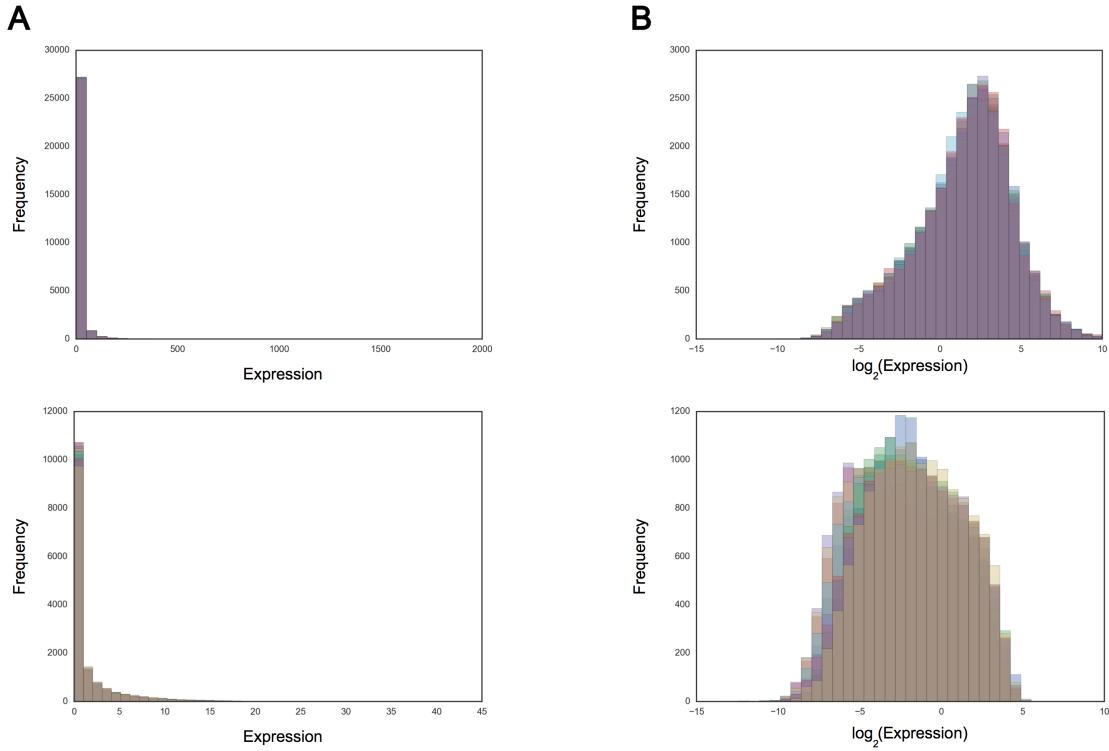
### 5.3 Data retrieval

Mouse raw expression values were obtained using a Ruby script provided by the Allen Brain Institute under the Apache 2.0 License. This script utilized the RESTful Model Access (RMA) service to download approximately 2,100 expression profiles. Each expression profile was produced using image analysis algorithms to summarize ISH stains of tissue sections. A full expression profile of a gene in the mouse dataset contained 77 values describing expression over 11 regions and 7 timepoints.

Human expression values for approximately 52,000 genes were obtained from downloading the Developmental Transcriptome Dataset located on <http://www.brainspan.org/static/download.html>. These expression values were compiled using quantization of RNA-seq results. The raw human data contained over 500 region-timepoint combinations. Of these, 10 common regions and 15 common timepoints were identified using a search algorithm to maximize the number of shared regions and timepoints.

### 5.4 Data pre-processing

Mouse and human datasets were first compared and filtered to 1,912 orthologs (shared genes). Raw expression values were further transformed by taking the base 2 logarithm, resulting in a more symmetrical distribution. This transformation was performed because the raw distribution was found to have high right skew (Figure 15). The data were also normalized to the interval [-1,1] using a proportional transformation. This was done due to the nature of the neural network architecture, which uses a hyperbolic tangent function on the output layer, resulting in a possible output interval of (-1,1).



**Figure 15:** Comparison between (A) raw expression values and (B) log<sub>2</sub>-transformed expression values for each region between the human (top) and mouse (bottom) datasets.

## 5.5 Machine learning model

A machine learning model was then constructed to reduce the dimensionality of each dataset to two dimensions. This program consisted of two main methods, a convolutional autoencoder (CAE) and a t-distributed stochastic neighbor embedding (t-SNE) that functioned sequentially.

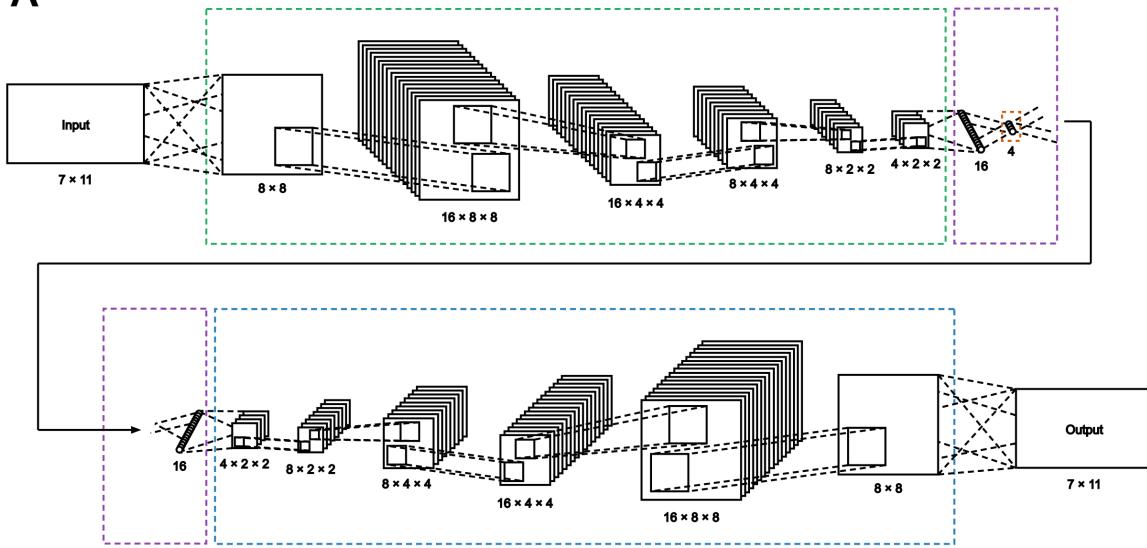
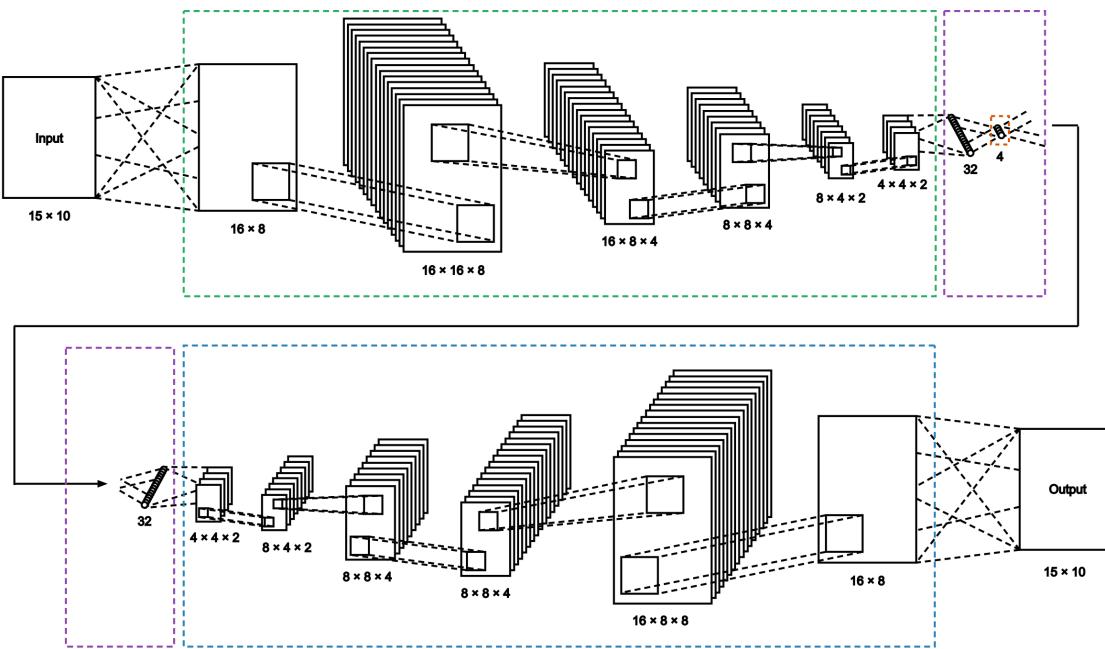
The CAE was composed mainly of convolutional encoding, dense hidden, and convolutional decoding layers (Figure 16). Because 2 by 2 max pooling layers require even dimensions to function correctly, the mouse expression profiles were first mapped using a dense network to a single 8 by 8 representation from the original 7 by 11 layer. Similarly, human expression profiles were mapped to a 16 by 8 representation from the original 15 by 10. The convolutional encoding layers then reduced each expression profile to four 2 by 2 (mouse) or 4 by 2 (human) filtered representations, which were transformed

into a 16- or 32-dimensional representation as the input to the dense hidden network. This representation was then mapped to a four-dimensional representation in the center of the neural network. The convolutional decoding layers functioned in the opposite manner of the convolutional encoding layers. For example, max pooling layers in the encoder were replaced with upsampling (interpolation) layers in the decoder.

The CAE was then trained on each dataset individually with the cost function being the root mean squared error (RMSE) between the two groups given by

$$\sqrt{\frac{\sum_i^n (\hat{y}_i - y_i)^2}{n}}$$

where  $n$  is the number of individuals,  $y_i$  is the dependent variable in a regression, and  $\hat{y}_i$  is the prediction output of the model.

**A****B**

**Figure 16:** CAE architectures used for dimensionality reduction of the mouse (**A**) and human (**B**) datasets. Green sections indicate convolutional layers involved in the encoding process, and purple sections indicate the encoded dense (fully connected) network. Blue sections indicate layers involved in the decoding process. The layer indicated in orange is the four-dimensional layer that is extracted upon training. The dimensions of each layer are shown below the respective layer.

The four-dimensional encodings obtained from the CAE were then reduced to two

dimensions using t-SNE, which attempts to retain high-dimensional features in a lower dimensions using probabilistic methods. According to Maaten and Hinton (2008), t-SNE first computes the distance between points in a high-dimensional space by computing a conditional probability  $p_{j|i}$  such that close points have a high probability whereas highly separated points have a near-zero probability. This probability is given by

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2/2\sigma_i^2)}.$$

with  $\sigma_i$  equal to the variance of an  $x_i$ -centered normal distribution.  $\|x\|$  represents the second norm of vector  $x$ , or the Euclidean distance from the origin. The similarity  $q_{j|i}$  between lower-dimensional points  $y_i$  and  $y_j$  obtained from applying a transformation to the original  $x_i$  and  $x_j$  is given instead by a Student's t distribution with a single degree of freedom such that

$$q_{j|i} = \frac{\exp(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} \exp(1 + \|y_k - y_i\|^2)^{-1}}.$$

Similar to the objective function of a neural network, t-SNE seeks to iteratively minimize a cost function given by the difference between the original  $p_{j|i}$  and the computed  $q_{j|i}$ . This is computed using the Kullback-Leibler divergence, which gives the difference between two probability distributions  $P_i$  and  $Q_i$  by computing the integral

$$\int_{-\infty}^{\infty} p_i(x) \log \frac{p_i(x)}{q_i(x)} dx$$

for continuous probability distributions (Bishop, 2006). In t-SNE, the objective function is computed by taking the sum of all Kullback-Leibler divergences over all points  $i$  given by

$$C = \sum_i \sum_j p_{i|j} \log \frac{p_{i|j}}{q_{i|j}}.$$

Minimization of this cost function is also performed similar to the minimization of that

of a neural network, using gradient descent. This gradient is given by

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{i|j} - q_{i|j})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}$$

and is updated with

$$\gamma^{(t)} = \gamma^{(t-1)} + \eta \frac{\partial C}{\partial \gamma} + \alpha(t) (\gamma^{(t-1)} - \gamma^{(t-2)})$$

where  $\gamma^{(t)}$  is the solution at time  $t$ ,  $\eta$  is the learning rate, and  $\alpha(t)$  is the momentum at  $t$ .

## 5.6 Cluster identification

Once each dataset was reduced to two dimensions, the density-based spatial clustering of applications with noise (DBSCAN) algorithm was applied to label clusters of similar gene expression. Unlike partitioning-based clustering, which attempts to find a function that best separates classes of points, or hierarchical clustering, which iteratively builds larger and larger clusters, DBSCAN is instead based on detecting dense collections of points. DBSCAN has many benefits, the most notable of which are high outlier robustness and noise detection ability. In addition, unlike many other algorithms, there is no need to manually input  $k$ , the hyperparameter that denotes the number of desired clusters (Ester et al.).

## 5.7 Clustering comparison

Clusters were compared by computing the adjusted rand index (ARI) which provides an indication of the similarity between two clusterings while taking into account results explained by random variance. The ARI returns a value between 0 and 1, with 0 indicating a clustering that may be the result of pure chance and 1 indicating that the two clusterings are completely identical. The ARI is a symmetric measure, meaning

that the order of the two clusters does not matter in calculation. The ARI is computed by constructing a contingency matrix between two clusterings  $k$  and  $h$  with  $C$  and  $R$  elements, respectively, such that each entry  $n_{ij}$  denotes the intersection between cluster  $i$  of  $k$  and cluster  $j$  of  $h$  (Table 2).

**Table 2:** The contingency matrix used in computing the ARI between two clusterings  $k$  and  $h$ . (Zhang and Wong, 2010)

Cluster	$k_1$	$k_2$	$k_3$	$\cdots$	$k_C$	$\Sigma$
$h_1$	$n_{11}$	$n_{12}$	$n_{13}$	$\cdots$	$n_{1C}$	$N_{1\cdot}$
$h_2$	$n_{21}$	$n_{22}$	$n_{23}$	$\cdots$	$n_{2C}$	$N_{2\cdot}$
$h_3$	$n_{31}$	$n_{32}$	$n_{33}$	$\cdots$	$n_{3C}$	$N_{3\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$h_R$	$n_{R1}$	$n_{R2}$	$n_{R3}$	$\cdots$	$n_{RC}$	$N_{R\cdot}$
$\Sigma$	$N_{\cdot 1}$	$N_{\cdot 2}$	$N_{\cdot 3}$	$\cdots$	$N_{\cdot C}$	$N$

The ARI may be computed (Yeung and Ruzzo, 2001) through the sums

$$\frac{\sum_{i,j} \binom{n_{ij}}{2} - \frac{\left[ \sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2} \right]}{\binom{n}{2}}}{\frac{1}{2} \left[ \sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2} \right] - \frac{\left[ \sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2} \right]}{\binom{n}{2}}}$$

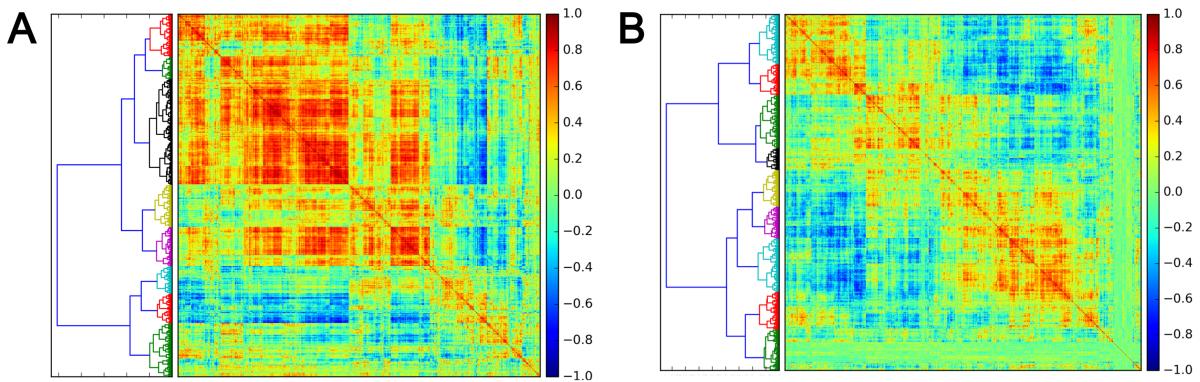
which are derived using the general hypergeometric distribution for significance.

## 6 Results

### 6.1 Correlation clustering

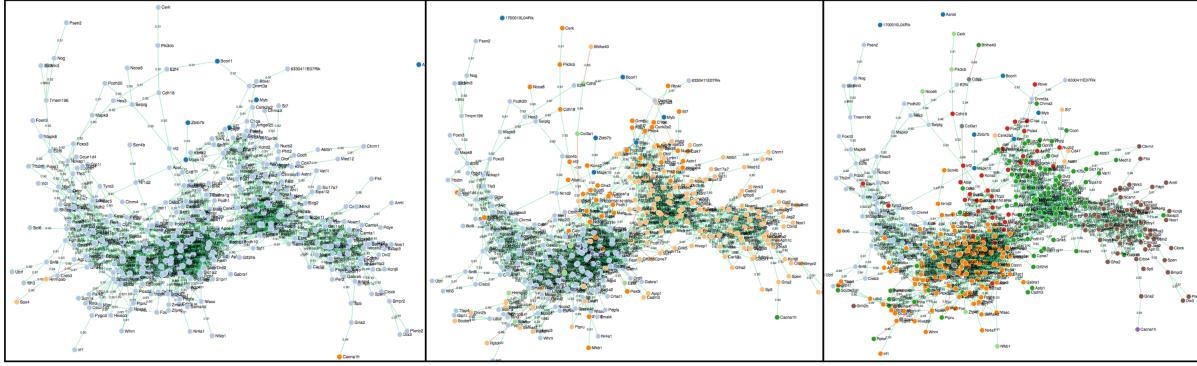
The mouse and the human datasets were initially clustered using Spearman's  $\rho$  measure of correlation to produce correlation matrices of size 1,912 by 1,912. These matrices were then re-indexed according to a clustering determined by application of a hierarchical process based on Ward's minimum variance method. The optimal number

of clusters was determined by attempting to minimize the silhouette coefficient of the resultant clusterings, which provides a rough estimate of the separation between clusters. The heatmaps of these correlation matrices are shown in Figure 17. Cooler colors represent more negative correlations and warmer colors represent more positive correlations between genes. The red diagonal line is due to each gene being perfectly positively correlated with itself. Comparison of these clusters using the ARI found a high divergence (0.0423) that was just above one expected from random chance.



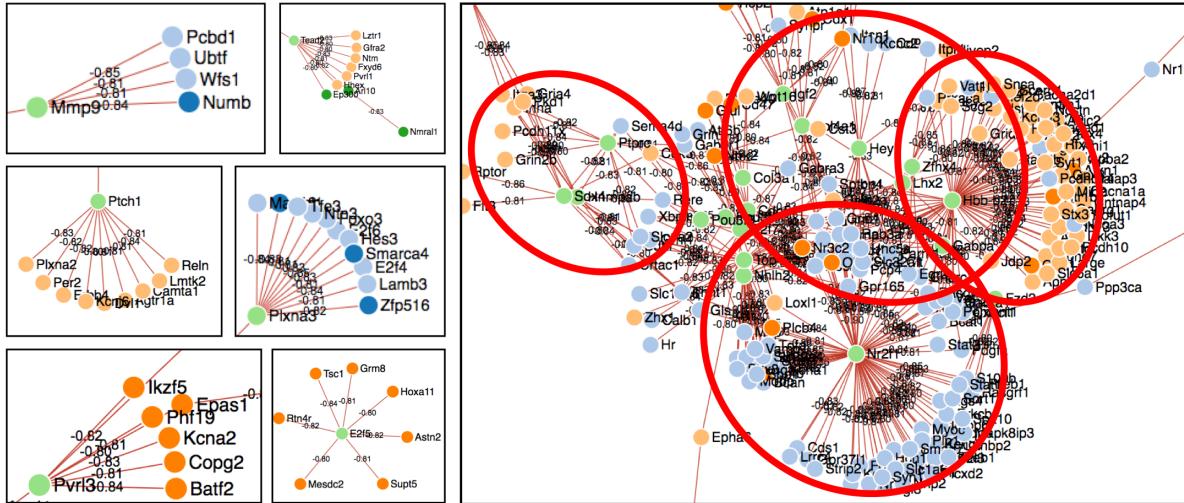
**Figure 17:** Correlation matrices produced by taking the Spearman  $\rho$  measures between genes in developing human (**A**) and mouse (**B**) brains. Genes were hierarchically clustered using the Ward distance measurement. Clusters shown in the dendograms to the right of each matrix are correlated with similarly-colored blocks in the matrices.

The correlation matrices were further used in the construction of graph networks in which nodes were represented by genes and edges were correlations between genes. To construct these networks, positive or negative correlations above or below a certain correlation threshold ( $\rho > 0.75$ ) and p-value ( $< 10^{-3}$ ) were drawn from these correlation matrices and used to assemble graphs. By superimposing cluster labels derived from the original hierarchical clustering (Figure 17), one could examine the general level of preservation between the two methods. Figure 18 shows one such graph obtained and colored using this method.



**Figure 18:** Construction of directed graphs using the Spearman correlation matrices preserves raw matrix clustering labels. In the above graphs, the nodes have been colored using labels derived from the dendrograms shown in Figure 17 with different levels of tree cutting.

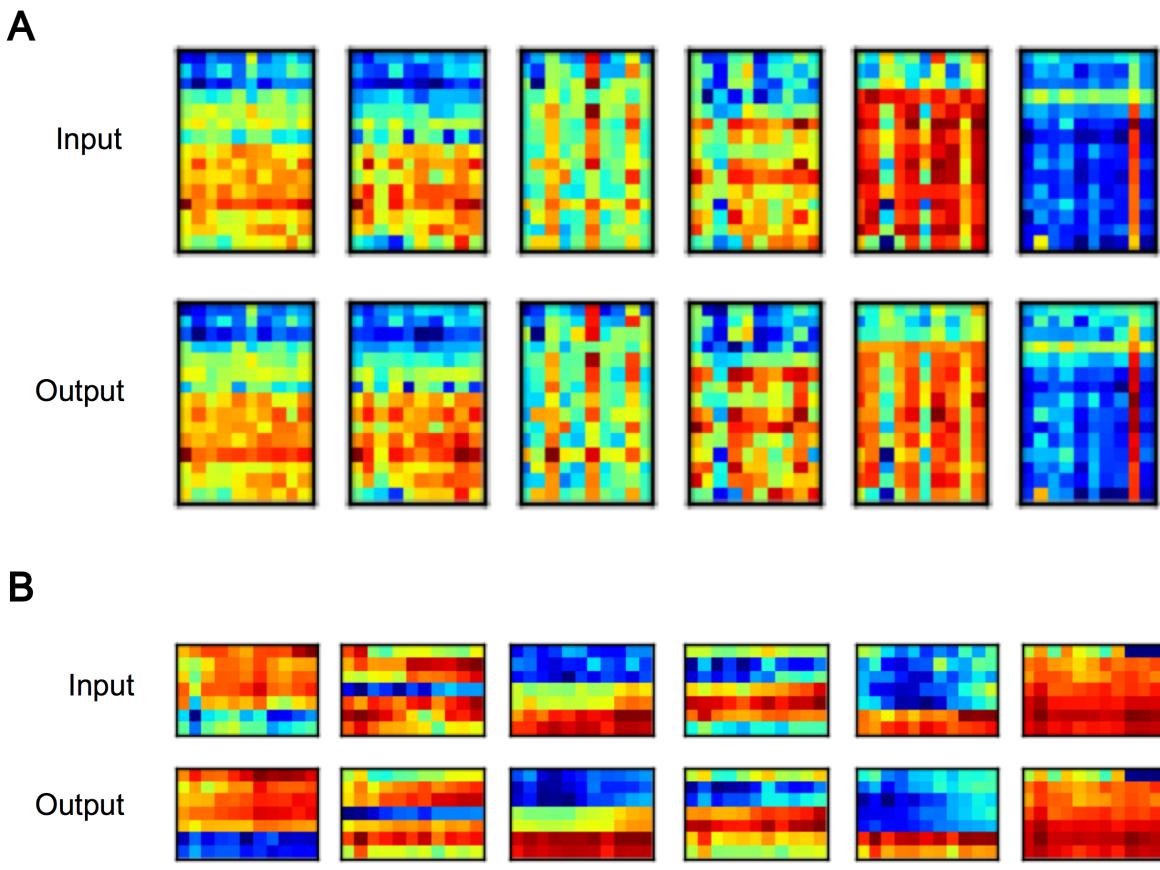
In addition, similar correlation networks were constructed which were composed exclusively of significant negatively-correlated gene expressions ( $(\rho > 0.75)$  and p-value ( $< 10^{-3}$ )). Analysis of these networks discovered the existence of several “hubs” of single genes negatively correlated with a multitude of other genes (Figure 19). These genes could be extracted from the raw correlation matrices through filtering for genes with multiple significant negative correlations. The functional significance of these hubs is that a high proportion ( $> 0.9$ ) of these hubs were found to produce regulatory proteins. The similarity between the human and mouse hubs was measured by extracting the top 200 such genes from each correlation matrix and counting the number of genes in common, which was found to be 40 genes. The statistical significance of this intersection was determined by conducting a Monte Carlo simulation consisting of 1,000,000 intersections of random samples of 200 genes, of which the average intersection was approximately 20 genes with a standard deviation of 4 in a normal distribution. Therefore, the p-value of such an intersection was less than  $10^{-6}$  despite being quite low (20%) in proportion.



**Figure 19:** Directed graphs constructed using only high negative correlations reveal "hubs" of negative regulator genes. Nodes are colored by hierarchically-derived labels, showing preservation of structure in the graph similar to that of Figure 18.

## 6.2 CAE-based encoding

In order to better examine the differences between these two organisms' expression landscapes, a robust machine learning approach was implemented to facilitate the visualization of each dataset on a lower-dimensional scale. A CAE was constructed to initially reduce the dimensionality of each dataset to 4 dimensions. The constructed CAE was found to have high reconstruction accuracy (Figure 20) in both datasets. In both datasets, the CAE was trained until an automatic halting program detected that no significant gains in accuracy were made after a certain timepoint.

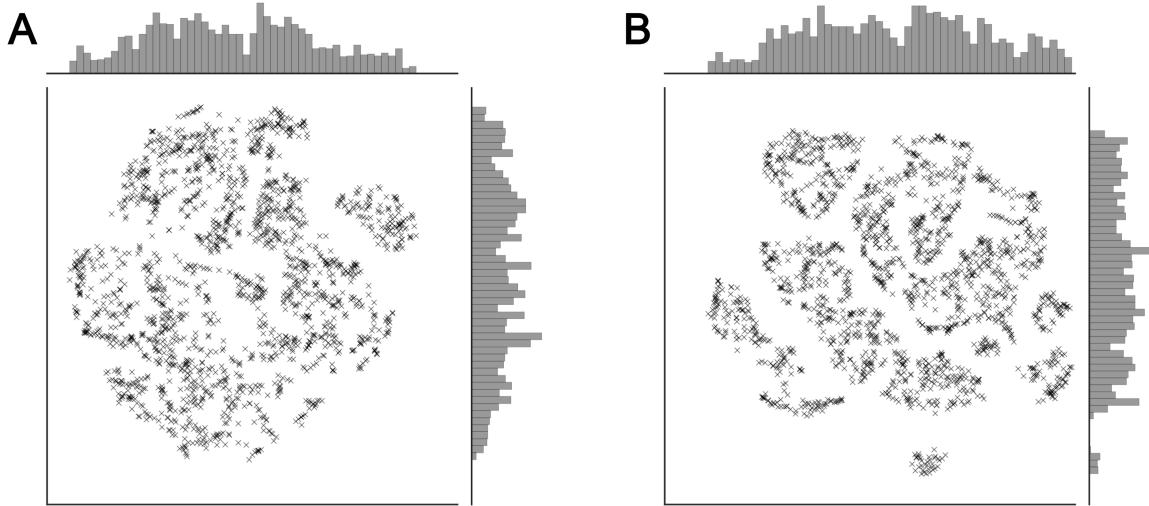


**Figure 20:** Comparison between CAE inputs and outputs for a random sample of human (**A**) and mouse (**B**) expression matrices. The average accuracies were found to be 28.49% for human expression profiles and 30.23% for mouse profiles, which are relatively high considering the high dimensionality of the data relative to the embedded dimensionality.

### 6.3 t-SNE

The four-dimensional CAE-derived representations were then reduced to two dimensions using t-SNE. Reducing the representations to 2 dimensions allowed for manual assessment of the expression landscapes in each dataset as well as visualization of any potential clusters (Figure 21). The mouse dataset was found to contain more distinct clusters in the two-dimensional representation compared to the human dataset, which may be due to the fact that the large number of dimensions in the original human representations had a negative impact on the performance of the CAE. However, this notion

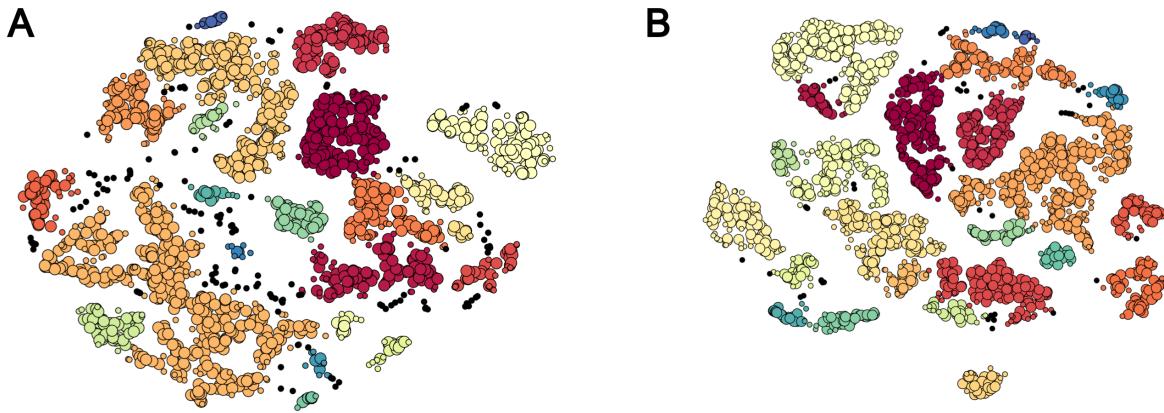
is contradicted by the similar accuracies of the CAE at recreating the expression profiles.



**Figure 21:** Two-dimensional representations returned by t-SNE for the human (**A**) and mouse (**B**) datasets.

## 6.4 Low-dimensional clustering

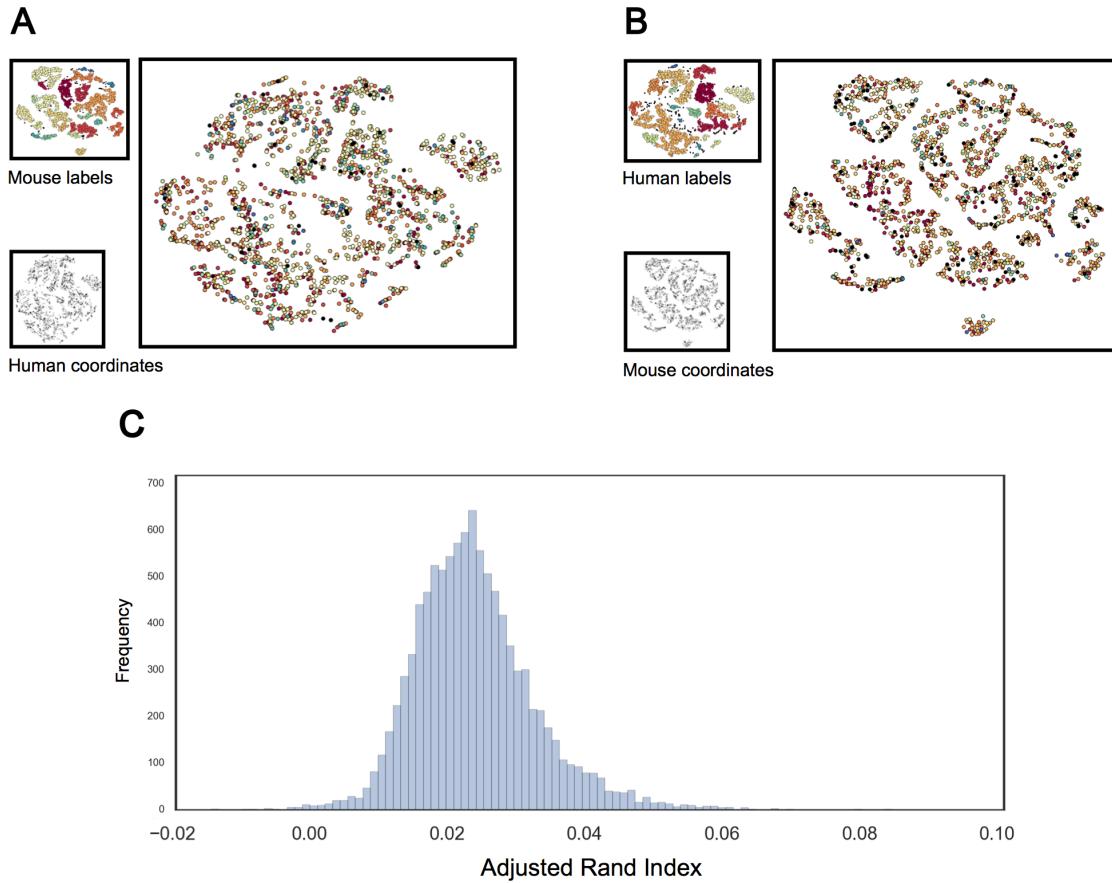
2-dimensional representations of the expression matrices were clustered using the DBSCAN algorithm, which detects dense collections of points. The Epsilon hyperparameter was set to 0.13 for both human and mouse datasets, and the minimum sample hyperparameter was set to 8 in the human dataset and 10 in the mouse dataset. These values were determined by attempting to minimize the Calinski-Harabaz index of the resultant clustering. Figure 22 displays an example of the clustering results.



**Figure 22:** Coloring/labeling of clusters using DBSCAN. Hyperparameters were determined by automatic minimization of the Calinski-Harabaz index, which computes the quality of a clustering based on how well-defined clusters are as a ratio between the intra- and inter-cluster dispersion means.

## 6.5 Low-dimensional clustering comparison

Initial comparison of the correlation matrix-derived clusters found a significantly low adjusted rand index value of 0.03853, which suggests a high divergence between the mouse and human expression sets. This value is similar to that obtained from comparing the clusters derived from the correlation matrices (Figure 17). Because both t-SNE and the CAE are stochastic processes, the 2-dimensional embeddings, and thus the clusterings, were characterized multiple times. By generating 100 clusterings in each dataset, 10,000 ARI value combinations were computed to provide the general distribution of similarity between the two datasets (Figure 23). In addition, the general similarity between two embeddings could be visually assessed by superimposing DBSCAN-generated labels from one organism onto the 2-dimensional coordinates of another (Figure 17). This method further shows the dissimilarity between the two organisms' gene expression landscapes.



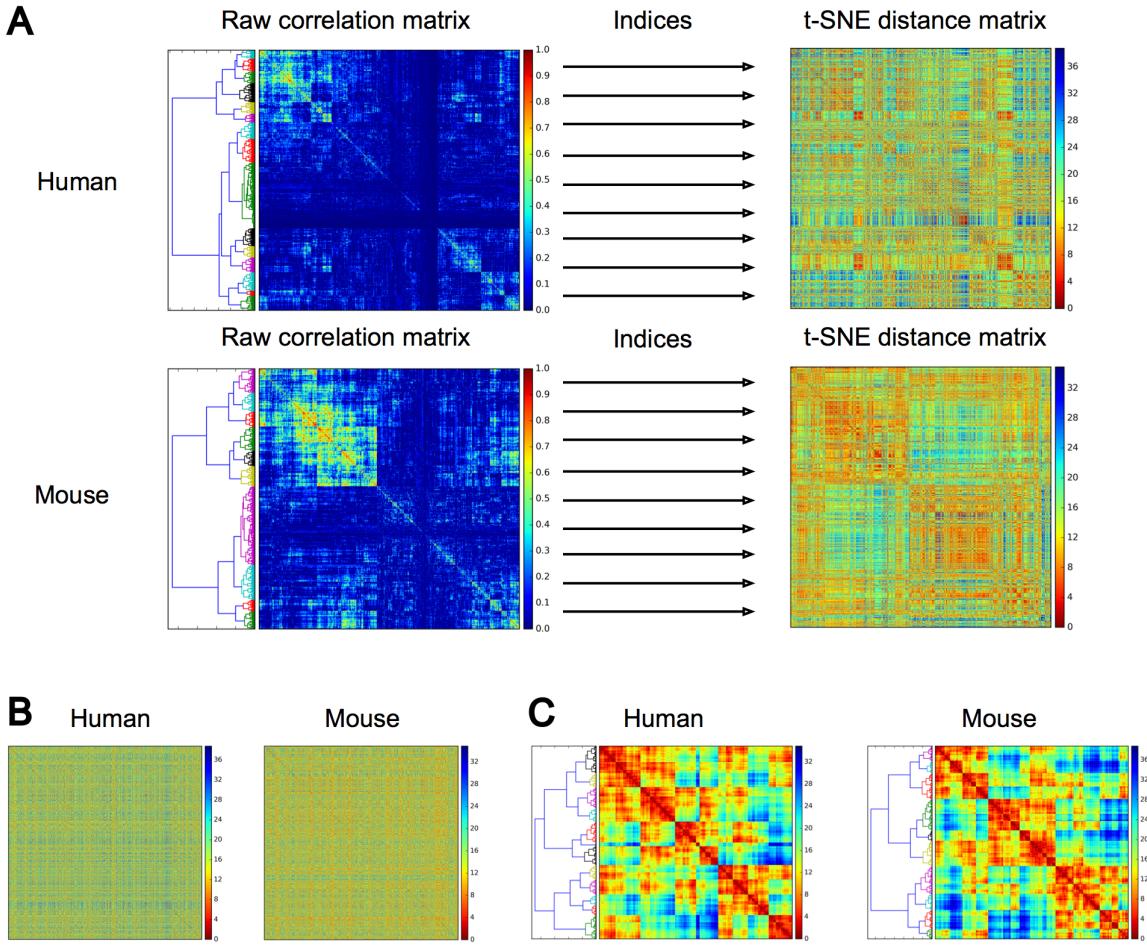
**Figure 23:** Distributions of ARI values computed by comparing 100 mouse and human clusterings for a total of 10,000 values. Cross-labeling of scatterplots, shown in **A** (human) and **B** (mouse) further show the divergence between the two organisms. **C:** Distribution of ARI values between mouse and human clusterings suggests high divergence with a mean of only 0.0283.

## 7 Discussion

### 7.1 Clustering validity

Several tests were conducted in order to validate the results mentioned in the previous section. In the first test, indices produced from rearranging the  $\rho^2$  correlation matrix were used to rearrange the Euclidean distance matrix between points in the t-SNE-derived representation. Figure 24 displays the similarity evident between these two matrix types along with a random indexing for reference. It was further found that

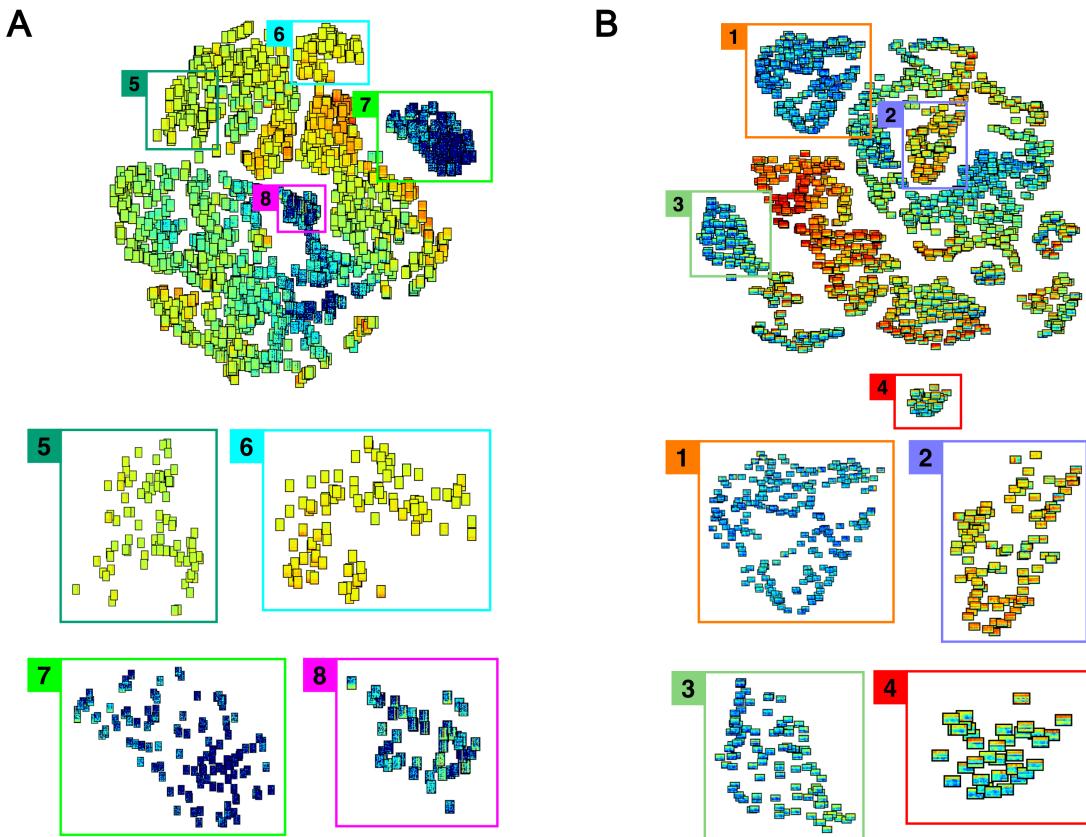
a block-based clustering of the t-SNE Euclidean distance matrix yielded more discrete clusters in both the mouse and human instances.



**Figure 24:** Euclidean distances among t-SNE representations reflect hierarchical clustering-derived distances. **A:** Using the arrangement produced by Spearman clustering, genes in the Spearman matrices (left) were replaced with Euclidean distances of the same genes in the 2D t-SNE representation. The moderate degree of arrangement in the resultant t-SNE distance matrix (right) is evidenced by the presence of discrete color blocks in the matrix, which are especially apparent when compared to a random ordering of the matrix (**B**). **C:** Ward distance-based clustering of the t-SNE distance matrix reveals discrete clusters more evident than those in the Spearman matrix.

The ability of the t-SNE representation to preserve the original structure of the data is further evident when points in the representation are replaced with the original gene expression heatmaps, as seen in Figure 25. In these two figures, the expression

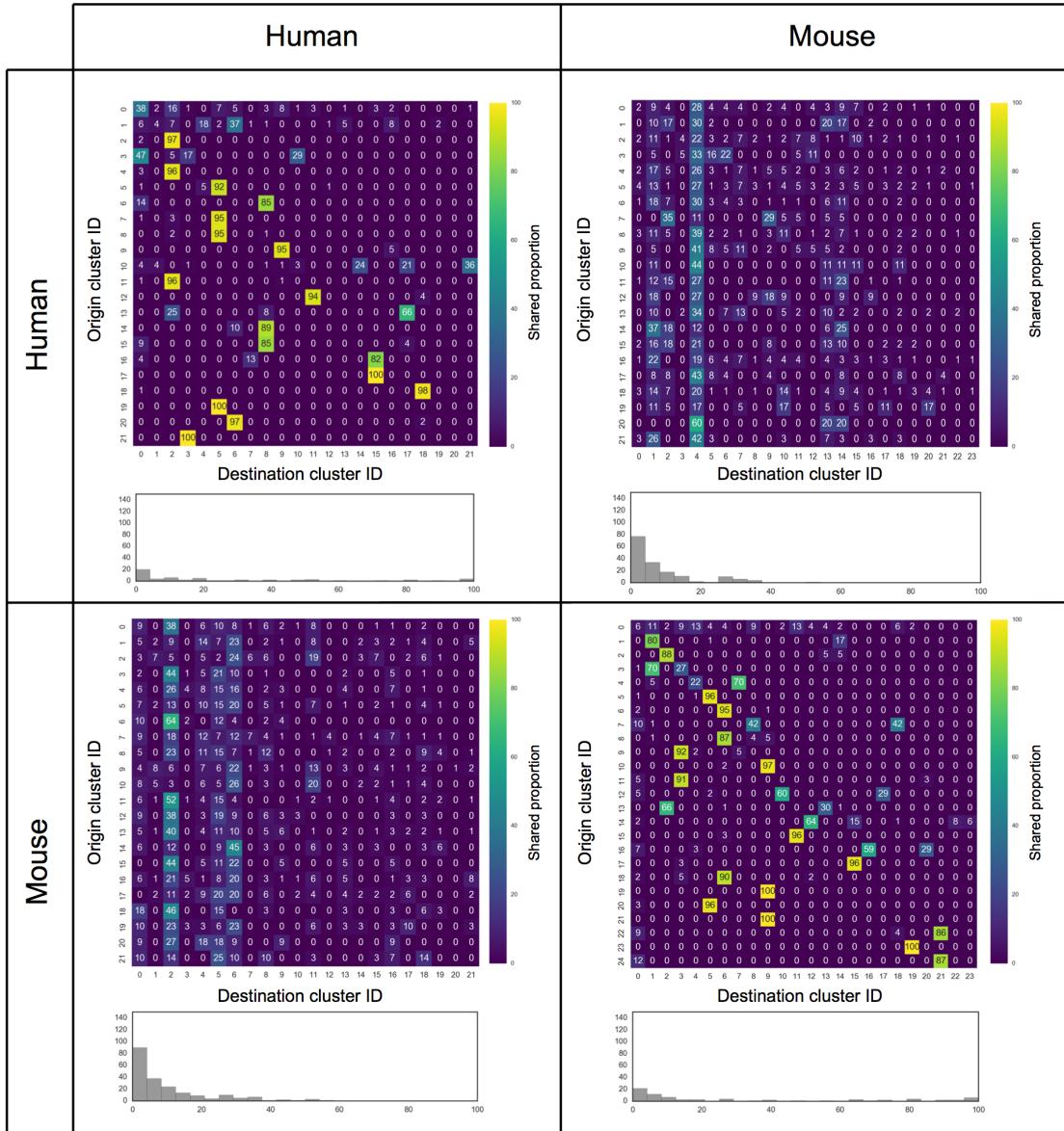
matrices corresponding to each point in the 2D representation have been overlaid such that the low-dimensional location of each matrix may be examined. This visualization helps support the fact that the machine learning model was able to accurately project the expression matrices onto these representations.



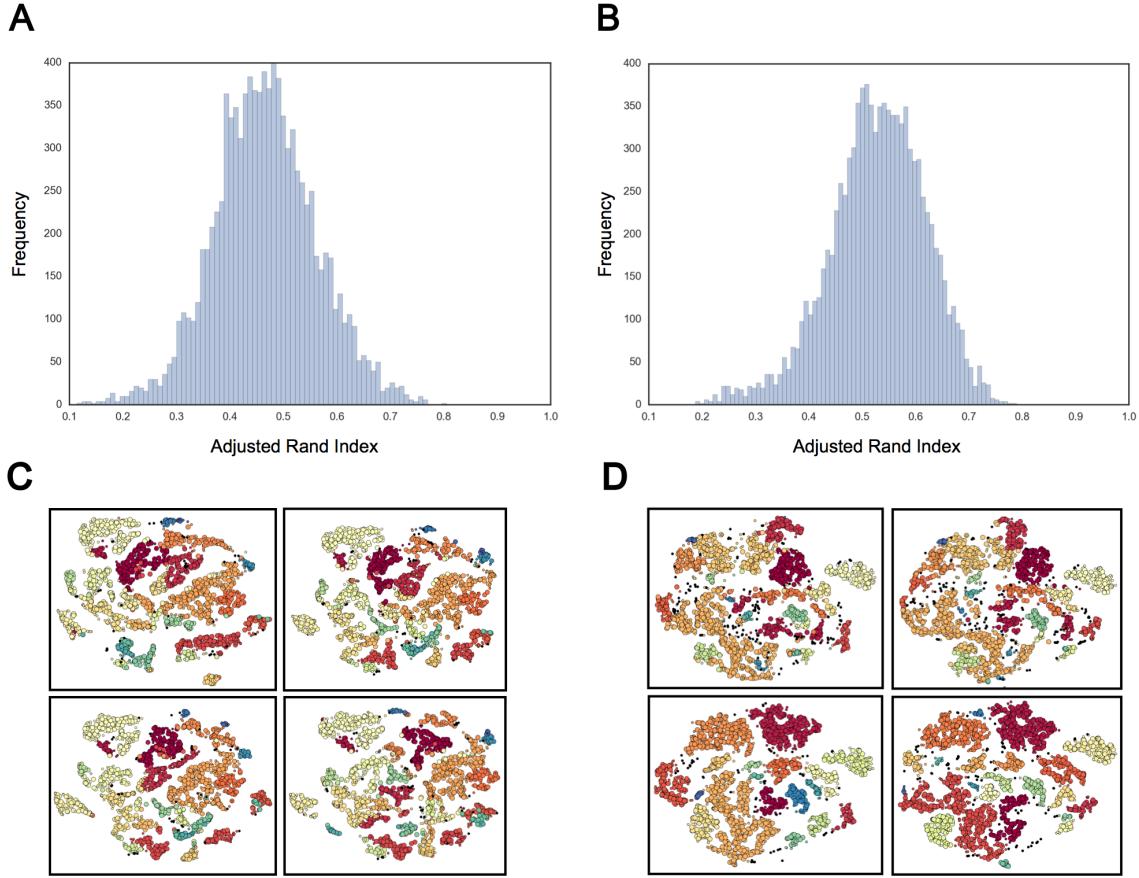
**Figure 25:** Representation of each point in the 2D embeddings of human (**A**) and mouse (**B**) expression matrices as the original image. Closer examination reveals that local similarities are preserved.

Another potential concern of the clustering is that the dimensionality reduction process accentuates differences between expression profiles at random in order to yield the clusters present in the two-dimensional representation. This concern is increased by the fact that the convolutional autoencoder and t-SNE are stochastic processes. To assess the consistency of the 2D clustering process, 100 different 2D representations were constructed in both the mouse and the human datasets and inter-organism and intra-organism ARI values were calculated. The distribution of ARI values between clusterings of the same

organism was also examined using this method and was found to be centered around 0.4 to 0.5, which is relatively high. Furthermore, homogeneity tables displaying the proportional composition of clusters between two clusterings were constructed (Figure 26). From these tables, it is evident that clusterings between the same organism are highly similar, whereas those between organisms are not.



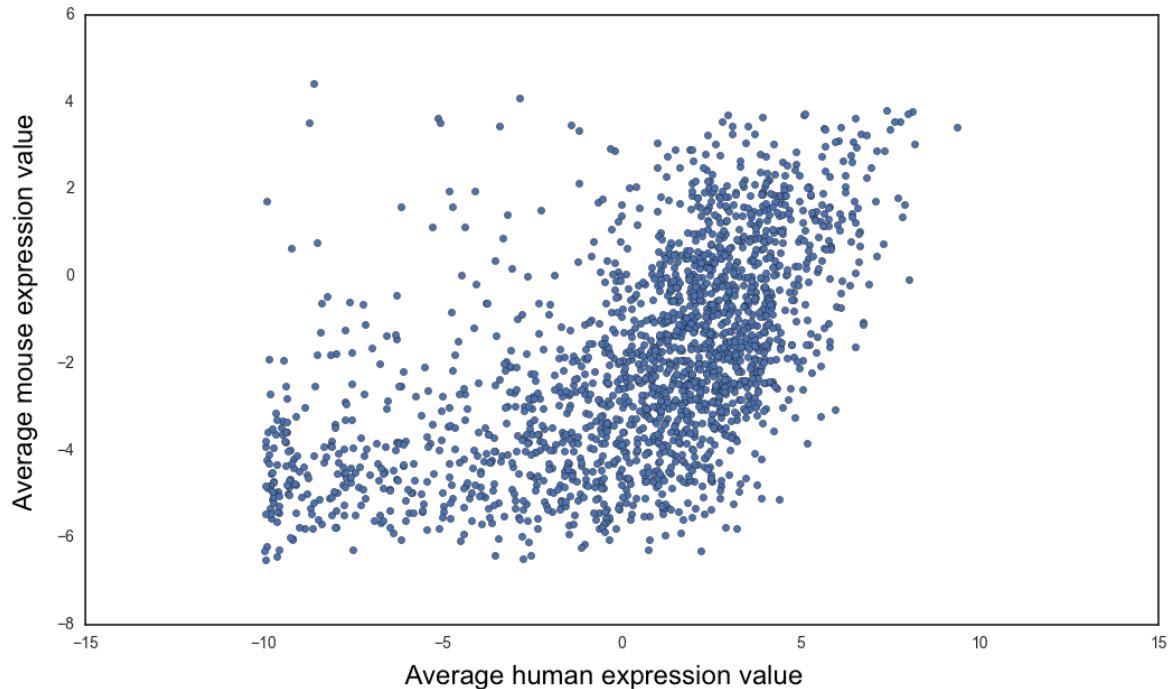
**Figure 26:** Homogeneity tables for human-human (top left), human-mouse (top right), mouse-human (bottom left), and mouse-mouse (bottom right) clusterings. Each cell in the table represents the proportion of a cluster labeled on the left axis represented in a cluster labeled on the bottom axis. An ideal (identical) clustering would consist of only 100% and 0% values. The consistency of human-human and mouse-mouse clusterings are shown in their respective homogeneity matrices, while the high divergence between mouse and human clusterings is shown in their respective matrices. The histogram below each figure represents the distribution of non-zero values in the homogeneity matrix. Homogeneity matrices between organisms were discovered to have a significantly lower proportion of low non-zero values and had values up to the 95-100% range.



**Figure 27:** **A** and **B**: Histograms of ARI values between clusterings of the same organism for human and mouse datasets, respectively. **C** and **D** show cross-colored scatterplots using the labels assigned from Figure 22.

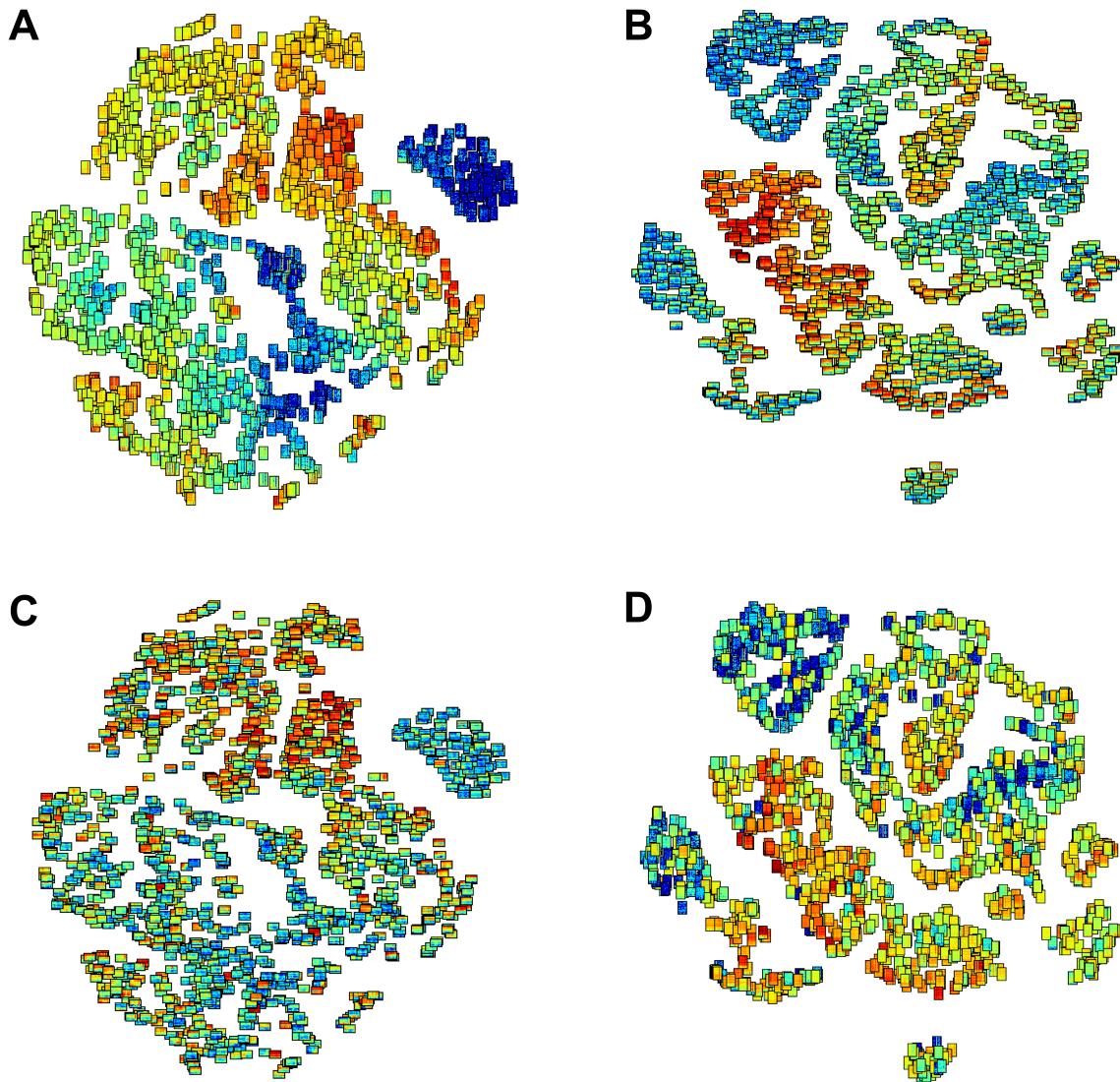
## 7.2 ARI analysis

The high divergence between the mouse and human expression landscapes is evidenced at several levels such as the correlation and DBSCAN clusterings as well as the divergence between hubs in the negative correlation networks. However, global expression levels appear to display moderate correlation (Figure 28). Global expression values were determined by taking the arithmetic mean of each expression matrix in the human and mouse datasets.



**Figure 28:** On an overall (global) level, human-mouse expression levels show a moderately strong ( $\rho = 0.613$ ) and highly significant ( $p = 2.3 * 10^{-199}$ ) correlation. This correlation, however, contradicts the ARI values in both the DBSCAN- and correlation matrix-derived clusterings.

The divergence between the two organisms may be further visualized by cross-overlaying the original expression matrices of each organism in a manner similar to that of Figure 25. In Figure 29, sections **C** and **D** have been constructed using the t-SNE-derived coordinates of the alternate organism. Although global similarity is somewhat evident in these figures, it is also apparent that specific differences in region and time separate the expression landscapes of humans and mice.



**Figure 29:** Using the visualization method from Figure 25, visualization of global cross-organism expression relationships may be accomplished by replacement of each point in the t-SNE scatter plot with corresponding expression matrices from the other organism. In **A** and **B**, each coordinate has been replaced with the corresponding expression matrix from the same organism in the same manner as figure 25, with **A** being of human expression and **B** of mouse expression. In **C** and **D**, each expression matrix is from the other organism. In **C**, the human t-SNE coordinates have been used to arrange mouse expression matrix, and vice-versa in **D**. Comparison of **C** against **A** and **D** against **B** reveals that gene expression is shared at the global level (overall expression of genes is shared, evident through the color similarities) but not at the local level.

## 8 Conclusions

The results of this study support the initial hypothesis that significant differences exist between mouse and human gene expression landscapes. In this study, a high degree of divergence between mouse and human gene expression profiles was observed when comparing correlation network graphs, correlation matrix-derived clusters, and clusters characterized using a dimensionality reduction model. Furthermore, DNA-binding proteins enriched in distinct areas of the negative correlation graphs were found to have a low proportion of commonality between human networks and mouse networks. The clustering characterization methods were shown to be robust through numerous validation measures such as examination of subsampling-derived correlation matrices and networks, consistency in multiple trials of the dimensionality reducing process, and manual examination of overlaid scatterplots of the 2D representations.

The methods and results developed in this study have numerous applications to computational and biomedical research. The clustering comparison scheme may be applied to other disease models such as Rhesus monkeys, other transgenic mouse models, and cerebral organoids. The transcription factor differences identified in the correlation networks may further be used in the development of drugs treating genetic diseases and in the examination of mouse cancer models, in which DNA-binding proteins play an integral role. Furthermore, the visualization methods produced by this study may be applied for any high-dimensional transcriptomic data such as in the field single-cell sequencing, which is rapidly becoming pertinent in cancer research. To help facilitate the use of the tools developed in this project, the complete set of all code, data, and figures generated by this study is available on GitHub at <https://github.com/kevin-who/braindev-analysis>.

## References

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving,

- G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.
- Angerer, L. M. and Angerer, R. C. (1991). In Situ Hybridization—A Guided Tour. *Toxicology Methods*, 1(1):2–29.
- Bakken, T. E., Miller, J. A., Luo, R., Bernard, A., Bennett, J. L., Lee, C.-K., Bertagnolli, D., Parikshak, N. N., Smith, K. A., Sunkin, S. M., Amaral, D. G., Geschwind, D. H., and Lein, E. S. (2015). Spatiotemporal dynamics of the postnatal developing primate brain transcriptome. *Hum. Mol. Genet.*, 24(15):4327–4339.
- Bisceglia, N. (2010). *Regulation of gene expression*. Nature Education.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- Burns, T. C., Li, M. D., Mehta, S., Awad, A. J., and Morgan, A. A. (2015). Mouse models rarely mimic the transcriptome of human neurodegenerative diseases: A systematic bioinformatics-based critique of preclinical models. *European Journal of Pharmacology*, 759:101–117.
- D'haeseleer, P. (2005). How does gene expression clustering work? *Nat Biotech*, 23(12):1499–1501.
- Ester, M., Kriegel, H.-P., and Jörg Sander and Xiaowei Xu, booktitle=KDD, y. A density-based algorithm for discovering clusters in large spatial databases with noise.
- Imaizumi, Y. and Okano, H. (2014). Modeling human neurological disorders with induced pluripotent stem cells. *Journal of neurochemistry*, 129 3:388–99.

Koushik, J. (2016). Understanding convolutional neural networks. *CoRR*, abs/1605.09081.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *NIPS*.

Lancaster, M. A. and Knoblich, J. A. (2014). Generation of cerebral organoids from human pluripotent stem cells. *Nat Protoc*, 9(10):2329–2340.

Lin, S., Lin, Y., Nery, J. R., Urich, M. A., Breschi, A., Davis, C. A., Dobin, A., Zaleski, C., Beer, M. A., Chapman, W. C., Gingeras, T. R., Ecker, J. R., and Snyder, M. P. (2014). Comparison of the transcriptional landscapes between human and mouse tissues. *Proceedings of the National Academy of Sciences*, 111(48):17224–17229.

Maaten, L. V. D. and Hinton, G. E. (2008). Visualizing data using t-sne.

Miller, J., Ding, S., Sunkin, S., Smith, K., Ng, L., Szafer, A., Ebbert, A., Riley, Z., Royall, J., Aiona, K., Arnold, J., Bennet, C., Bertagnolli, D., Brouner, K., Butler, S., Caldejon, S., Carey, A., Cuhaciyan, C., Dalley, R., Dee, N., Dolbeare, T., Facer, B., Feng, D., Fliss, T., Gee, G., Goldy, J., Gourley, L., Gregor, B., Gu, G., Howard, R., Jochim, J., Kuan, C., Lau, C., Lee, C., Lee, F., Lemon, T., Lesnar, P., McMurray, B., Mastan, N., Mosqueda, N., Naluai-Cecchini, T., Ngo, N., Nyhus, J., Oldre, A., Olson, E., Parente, J., Parker, P., Parry, S., Stevens, A., Pletikos, M., Reding, M., Roll, K., Sandman, D., Sarreal, M., Shapouri, S., Shapovalova, N., Shen, E., Sjoquist, N., Slaughterbeck, C., Smith, M., Sodt, A., Williams, D., Züllei, L., Fischl, B., Gerstein, M., Geschwind, D., Glass, I., Hawrylycz, M., Hevner, R., Huang, H., Jones, A., Knowles, J., Levitt, P., Phillips, J., Sestan, N., Wohnoutka, P., Dang, C., Bernard, A., Hohmann, J., and Lein, E. (2014). Transcriptional landscape of the prenatal human brain. *Nature*, 508:199–206.

Monaco, G., van Dam, S., Ribeiro, J. L. C. N., Larbi, A., and de Magalhães, J. P. (2015).

A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue pathway and disease levels. *BMC Evol Biol*, 15(1).

Muller-Tidow, C., Metzelder, S. K., Buerger, H., Packeisen, J., Ganser, A., Heil, G., Kugler, K., Adiguzel, G., Schwable, J., Steffen, B., Ludwig, W.-D., Heinecke, A., Buchner, T., Berdel, W. E., and Serve, H. (2004). Expression of the p14arf tumor suppressor predicts survival in acute myeloid leukemia. *Leukemia*, 18(4):720–726.

Nguyen, L., Wang, Y., and Nikolakopoulou, A. (2015). *Cerebral organoid derived from ALS patient stem cells*. University of Southern California.

Peemen, M., Mesman, B., and Corporaal, H. (2011). Speed Sign Detection and Recognition by Convolutional Neural Networks. *International Automotive Congress*.

Simões-Costa, M. and Bronner, M. E. (2015). Establishing neural crest identity: a gene regulatory recipe. *Development*, 142(2):242–257.

Strand, A. D., Aragaki, A. K., Baquet, Z. C., Hodges, A., Cunningham, P., Holmans, P., Jones, K. R., Jones, L., Kooperberg, C., and Olson, J. M. (2007). Conservation of regional gene expression in mouse and human brain. *PLoS Genetics*, 3:1746 – 1758.

Thompson, C. L., Ng, L., Menon, V., Martinez, S., Lee, C.-K., Glattfelder, K., Sunkin, S. M., Henry, A., Lau, C., Dang, C., Garcia-Lopez, R., Martinez-Ferre, A., Pombero, A., Rubenstein, J. L., Wakeman, W. B., Hohmann, J., Dee, N., Sodt, A. J., Young, R., Smith, K., Nguyen, T.-N., Kidney, J., Kuan, L., Jeromin, A., Kaykas, A., Miller, J., Page, D., Orta, G., Bernard, A., Riley, Z., Smith, S., Wohnoutka, P., Hawrylycz, M. J., Puelles, L., and Jones, A. R. (2014). A High-Resolution Spatiotemporal Atlas of Gene Expression of the Developing Mouse Brain. *Neuron*, 83(2):309–323.

van der Maaten, L. (2014). Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, (15):1–21.

- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.
- Yang, J., Parikh, D., and Batra, D. (2016). Joint Unsupervised Learning of Deep Representations and Image Clusters. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yeung, K. Y. and Ruzzo, W. L. (2001). Details of the adjusted rand index and clustering algorithms supplement to the paper ” an empirical study on principal component analysis for clustering gene expression data ” (to appear in bioinformatics).
- Zhang, S. and Wong, H.-S. (2010). Arimp: A generalized adjusted rand index for cluster ensembles. In *ICPR*.
- Zilouchian, A. (2001). Fundamentals of neural networks.

## A Limitations and Assumptions

This project was limited by the number of genes that were analyzed as well as the amount of expression data per gene. A central assumption, therefore, was that there was a sufficient number of genes and expression data for the results of the clustering to be determined by intrinsic differences and not variability. This assumption is supported by the relatively large number of genes (1,912) and expression data per gene (77 dimensions for mouse data, 150 for human data). Furthermore, the mouse and human donors used in this project were assumed to be representative of their respective populations. The regions of expression matrices were also assumed to be constant throughout development, which is supported by the high-accuracy computational methods used by the Allen Brain Atlas to quantize expression values.