Feature engineering for low dimensional representation of genes' expression and pathological activities across diverse human tissues

A THESIS

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

SUBMITTED BY

PRIYADARSHINI RAI

SUPERVISED BY

Dr. Debarka Sengupta & Prof. Angshul Majumdar

Department of Computational Biology

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI - 110020

OCTOBER 2022

Thesis Certificate

This is to certify that the thesis titled **Feature engineering for low dimensional representation of genes' expression and pathological activities across diverse human tissues** is submitted by **Priyadarshini Rai** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of **Doctor of Philosophy**, is a bonafide record of the research work done by her under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Debarka Sengupta
Associate Professor
Indraprastha Institute of
Information Technology
Delhi - 110020

Dr. Angshul Majumdar
Professor
Indraprastha Institute of
Information Technology
Delhi - 110020

October 2022

Acknowledgments

Each individual who enters our life teaches us at least one beneficial lesson. It has been an honor to learn from my supervisors, **Dr. Debarka Sengupta** and **Dr. Angshul Majumdar**, over the last five and a half years. There are countless lessons I can pick up from them. But, the two main limitations of my personality will be overcome if I can strive to absorb one trait from each of their personas. The ease with which Dr. Debarka adjusts to everything new, like a river, and the lightning-quick efficiency with which Dr. Angshul completes any duty at his disposal. I shall be eternally grateful for the training I received from them during my Ph.D.

I would like to thank **Prof. GPS Raghava**, Head, Department of Computational Biology, IIIT Delhi, who has always extended his support whenever needed. I would also like to express my gratitude to **Dr. Gaurav Ahuja** for always being available whenever I needed his advice on one of the study issues. I also sincerely thank IIIT's founder Director, **Prof. Pankaj Jalote**, and IIIT Director, **Prof. Ranjan Bose**, for providing world-class research facilities.

I would also like to thank my annual review committee members, **Dr. Vibhor Kumar** and **Dr. Ganesh Bagler**, for their continuous feedback that helped me to improvise my research. I am also grateful to **all the professors in the Department of Computational Biology** for the knowledge they imparted through the courses they taught.

Let me also take this opportunity to thank the administrative staff, particularly **Ms. Priti Patel** and **Dr. Ravi Bhasin**, who helped me with many academic issues in time. In particular, I would like to thank **Mr. Adarsh Aggarwal** for always being available to help whenever computing assistance was required.

I had the good fortune to meet a family in **Neetesh** and a lifelong buddy in **Smriti** while pursuing my doctorate. Besides them, I got to spend some wonderful time with **Indra**, **Aanchal**, **Madhu**, **Chitrita**, **Sarita**, **Shreya**, **Vivek**, **Pooja**, **Jyoti**, **Raghav**, **Dilraj**, **Omkar**, **Krishan**, **Namrata**, **Sumeet**, **Shiju**, **Shruti**, **Piyush**, **Samridhi**, **Anjali**, **Atishay**, **Megha**, **Vanika**, **Shikha**, **Shalini**, **Shivani**, **Neha**, and **Divya**.

"Nine-tenths of education is encouragement," by Anatole France. I would like to thank my M.Tech thesis supervisor **Emeritus Prof. Madasu Hanmandlu** for believing in me and encouraging me to pursue Ph.D. He was Professor at IIT Delhi when I joined his lab to work on my M.tech dissertation. I was afraid that I wouldn't be able to deliver what was expected. Although surrounded by outstanding postgraduate students, his trust in me has given me immense confidence to pursue my Ph.D.

I would not have reached this stage where I am writing the acknowledgment of my Ph.D. dissertation without the support and encouragement of **my parents** (especially my mother) and my **husband Anant**. My parents are the foundation of my building, which is my education, while my husband is a shield who takes on everything for the purpose of safeguarding me. I also want to express my gratitude to my **sister Anita** and my **brother Ashutosh**, who sacrificed a great deal to help me finish my Ph.D. My **mother-in-law** deserves respect as well for consistently attempting to comprehend my viewpoint. At last, I would like to thank the ***SUN*** of our lives, our ***DAUGHTER*** **Anika**, and my **nephew Avyaan** for teaching me to go beyond my limitations.

Priyadarshini Rai

(*Priyadarshini Rai*)

Abstract

The advent of tissue and single cell based transcriptomic profiling technologies has allowed precise characterization of tissue specific gene activities in the context of development and disease. Human cells express about 20,000 genes whose interplay enables all physical activities that define our life. However, with expression signals, most transcriptomic platforms also offer bewildering levels of noise. This has become more prominent in the case of single cell transcriptomic experiments. As such, it is important to represent cells and tissues with the help of minimal genesets. This poses the classical challenge of dimension reduction. To reduce this feature space, we developed a *de novo* feature selection algorithm, SelfE (self expression), a novel $l_{2,0}$-minimization algorithm that determines an optimal subset of feature vectors (genes) that preserves subspace structures as observed in single cell RNA-sequencing data. We compared SelfE with the commonly used feature selection methods for single-cell expression data analysis.

Unlike bulk RNA sequencing data, single cell gene expression readouts feature excessive dropout events, thereby confounding downstream bioinformatic analyses. Keeping these limitations in mind, we proposed a method that employs deep dictionary learning for the clustering of single cell data. This is the first piece of the effort to create a deep learning-based approach for clustering. We render the framework clustering compatible by introducing a cluster-aware loss (K-means and sparse subspace) into the learning problem. The potential of our method is demonstrated by comparison with general deep learning-based clustering techniques and with specially designed single-cell RNA clustering techniques.

In an effort to provide a comprehensive resource to understand tissue specific pathological activities of the genes, we developed Pathomap. It allows querying a gene to visualize, on a human body template, the intensity of pathological activities of a certain gene in a tissue specific manner. While the Human Cell Atlas project is still consolidating the gene expression patterns across healthy human tissues, Pathomap, as a parallel, provides insights into tissue specific pathological activity of genes. To achieve this, we searched 18 million PubMed papers published through May 2019 and automatically selected 4.5 million abstracts describing certain genes' functions in disease development. In addition, we fine-tuned the pretrained Bidirectional Encoder Representations from Transformers (BERT) for text modeling from the field of Natural Language Processing (NLP) in order to learn embeddings of entities such as genes, diseases, tissues, cell types, etc., in a way that preserves their relationship in a vector space. The reprogrammed BERT predicted disease-gene relationships not present in the training data, demonstrating the viability of in-silico formulation of hypotheses relating to diverse biological entities such as genes and disorders.

Taken together, our works bring feature engineering approaches to bear in representing biological entities in low-dimensional space.

List of Publications

PUBLISHED

- **Rai, P.**, Sengupta, D. and Majumdar, A., 2020. SelfE: Gene Selection via Self-Expression for Single-Cell Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

- **Rai, P.**, Majumdar, A. and Sengupta, D., 2020, November. Cluster Aware Deep Dictionary Learning for Single Cell Analysis. In *International Conference on Neural Information Processing* (pp. 62-69). Springer, Cham.

- Chawla, S., Rockstroh, A., Lehman, M., Ratther, E., Jain, A., Anand, A., Gupta, A., Bhattacharya, N., Poonia, S., **Rai, P.**, Das, N., Majumdar, A., Jayadeva., Ahuja, G., Hollier, B.G., Nelson, C.C., and Sengupta D., 2022. Gene expression based inference of cancer drug sensitivity. *Nature communications*, *13*(1), pp.1-15.

PREPRINTS

- **Rai, P.**, Jain, A., Jha, N., Sharma, D., Kumar, S., Raj, A., Gupta, A., Poonia, S., Chawla, S., Majumdar, A. and Chakraborty, T., 2022. A visual atlas of genes tissue-specific pathological roles. *bioRxiv*.

- Poonia, S., Goel, A., Chawla, S., Bhattacharya, N., **Rai, P.**, Lee, Y.F., Yap, Y.S., West, J., Bhagat, A.A., Tayal, J., Mehta, A., Ahuja, G., Majumdar, A., Ramalingam, N., and Sengupta D., 2021. Marker-free characterization of single live circulating tumor cell full-length transcriptomes. *bioRxiv*.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

1.1.1 The central dogma

In molecular biology, a central dogma describes how information flows within a cell, from DNA to RNA to proteins [22]. Deoxyribonucleic acid (DNA), which has a popular double helix structure is used to store long-term data by cells. A series of nucleotide molecules connected by phosphate groups make up each double helix strand. A nitrogenous base, a deoxyribose sugar, and a phosphate group are the three components that make up these nucleotides. Adenine, cytosine, thymine, and guanine are the four species of DNA nucleotides, each of which has a unique base subunit. DNA strands are made up of sugars and phosphates and are connected by hydrogen bonds between nucleotides known as base pairs. Three hydrogen bonds are formed between guanine and cytosine, and two hydrogen bonds are formed between adenine and thymine.

When the cell wishes to use some of this information, it duplicates it as a molecule of ribonucleic acid (RNA) through a process known as transcription. Genes are sections of DNA that encode functional information. RNA is comparable to a single strand of DNA, with the exception that the deoxyribose sugar is replaced by ribose and the thymine nucleotide is replaced by uracil. RNA does not have a double helix structure since it is single-stranded, but it can form complicated forms by attaching to itself. The ability of a cell to produce multiple copies of the same RNA molecule is referred to as its expression level. By altering RNA expression, the activity of cellular processes can be modulated. There are numerous types of RNA that serve distinct functions. RNA molecules that are transcribed from genes that code for proteins are referred to as messenger RNA (mRNA). The other types of RNA are ribosomal RNA (rRNA), which is a component of the ribosome (the molecular machinery that synthesizes proteins), transfer RNA (tRNA), which transports amino acids to the ribosome, micro RNA (miRNA), which regulates gene expression, and long non-coding RNA (lncRNA), which is also involved in regulation and other processes. Exons, which code for information, are interspersed with introns, which are much larger non-coding regions.

When an mRNA molecule is initially transcribed, it contains intronic sequences. These sequences are removed by RNA splicing, and a sequence of adenine nucleotides (a poly(A) tail) is added to the 3' end of the mRNA molecule to mark it as mature. By selectively retaining or removing exons, this splicing process permits the production of multiple forms of a protein from a single gene. A complex structure composed of specialized RNA and proteins called the ribosome converts a mature mRNA transcript into a protein. Because the information encoded by nucleic acids in RNA is translated into information stored as amino acids in the protein, the conversion process is known as translation. For a cell to function, proteins perform most of the tasks. These functions require complex three-dimensional structures and comprise tasks such as regulating gene expression, constructing new proteins, sensing the external environment, transporting nutrients into the cell, recycling molecules, and promoting metabolism. Understanding the molecules involved in the central dogma is central to our understanding of how a cell works. Understanding how a cell function relies heavily on our knowledge of the molecules involved in the central dogma.

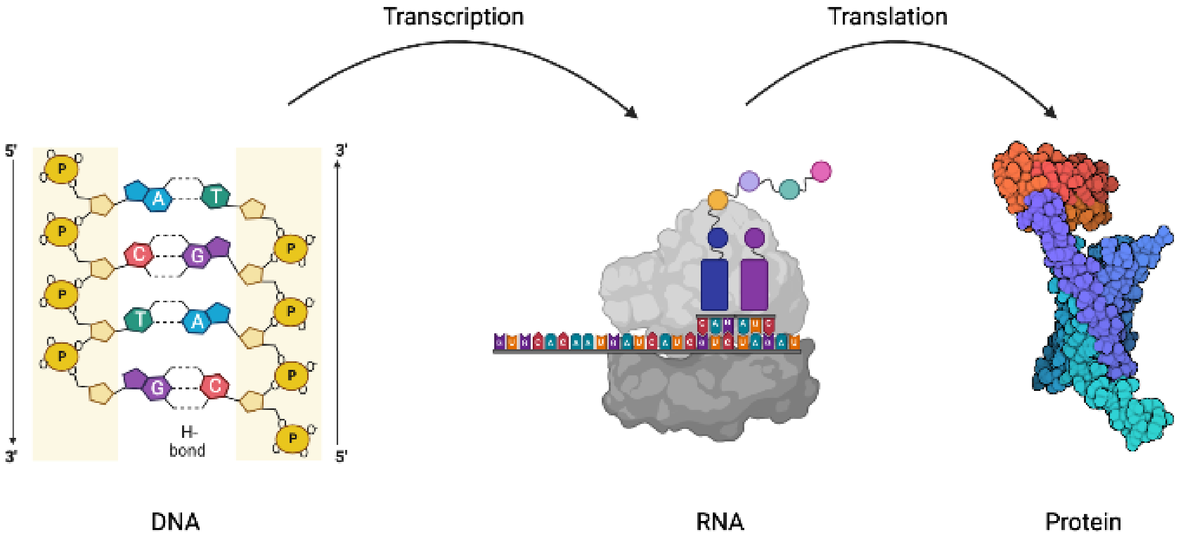

Figure 1.1: The above figure represents the process of conversion of DNA into mRNA, that is, transcription and conversion of mRNA to proteins, that is, translation.

1.1.2 Bulk and single cell expression data

The cell consists of DNA, and genes are the functional segments of DNA. For each cell, the transcripts originated per gene determine the expression of a gene for that particular cell. The variability in the transcript count for a gene determines the characteristics and functionality of the cell. By analyzing the gene expression variability, one can determine how a biological function gets influenced by the pattern of gene expression in a cell. Eventually, this can help us in recognizing genomic functions and the progression of a disease.

High throughput sequencing or massively parallel sequencing (MPS) technologies enable us to sequence thousands of samples parallelly [23]. These technologies have revolutionalized the way deoxyribonucleic acid (DNA)/ribonucleic acid (RNA) molecules used to be processed. The sequencing of RNA molecules, also known as RNA sequencing, allows us to study the transcriptome. During bulk RNA sequencing, the expression of a gene is estimated by taking an average of its expression level across many cells of a particular sample. But nowadays, with the help of next generation sequencing (NGS) technologies, one can profile the expression of a gene for a particular cell instead of recording the mean of the expression of a gene across hundreds to thousands of cells [24]. This procedure is known as single cell RNA sequencing (scRNA-seq), and the data generated through this procedure is known as single cell data.

The data generated with the help of RNA sequencing technologies are depicted in the form of a read count matrix where each row represents a gene, and each column represents a sample (also known as a replicate) from similar or different states. The replicates are of two types, namely biological and technical. A replicate is termed a biological replicate if it is taken from a population with a similar condition, but the individuals from whom the sample is derived are different. For example, parallel sample sequences are drawn from biologically different persons of the specific treatment groups. This helps us to understand the variability within the treatment group and infer conclusions about the ongoing treatment [25]. In the case of technical replicates, multiple sequencing libraries for a particular biological sample are generated to ensure the validity of the measured gene expression for the particular sample.

The bulk RNA sequencing profiles the average expression of genes across thousands to millions of cells. But, multicellular organisms depict cell to cell variability even within the same tissue type. Single cell RNA sequencing allows for studying multiple types of cells within a single tissue. It helps us to study the multispectral nature of genes and in classifying rare cell types. But, the limitation of single cell data is that it is sparse in nature due to the influence of various technical and biological variations during sequencing. Therefore, before using single cell data to perform any analysis, we need to process the data. The upcoming section discusses the steps involved in preprocessing single cell data. The analysis ready data is represented as a matrix where each row represents genes, and each column represents a discrete biological unit, that is, a cell.

1.2 Challenges of single cell data analysis and the importance of low dimensional representation of cells

1.2.1 Noise and bias in single cell data

Despite the rapid development of cell capture technologies and scRNA-seq protocols, the data they generate still presents a number of obstacles. Approximately 10 percent of a cell's transcripts are captured by existing methods [26]. Combined with the low sequencing depth per cell, this results in reduced sensitivity and the inability to detect transcripts with low levels of expression. In addition to contributing to high levels of technical noise, the small amount of starting material complicates downstream analysis and makes it difficult to detect biological differences [27]. Before cells can be captured, they must be dissociated into single-cell suspensions, a process that is not always simple. The treatments required to break apart some tissues or cell types may impact the cells' health and their transcriptional profiles. Using a cold-active protease to separate cells has been suggested as a method for minimizing damage caused by dissociation [28]. There may be other factors that prevent other cell types from being captured, such as their size or characteristics. Often, multiple cells are captured at once or empty wells or droplets are sequenced, which makes quality control of datasets an important factor.

Small amounts of starting material and low sequencing depth mean that there are many times when zero counts are recorded. This means that a particular gene in a particular cell was not measured to be expressed. These zero counts often show the real biological state we want to know about since we know that different types of cells will express different genes. Biological confounding factors, such as cell cycle stage, transcriptional bursting, and environmental interactions, which result in real changes in expression but may not be relevant to a particular study, can also cause zeros. These factors can also cause zeros as a result of genuine changes in expression. Additionally, there are outcomes that are solely technical aspects. Particularly, sampling effects can lead to "dropout" occurrences in which a transcript is actually expressed in a sample but is not detected in the sequencing data [29]. In bulk RNA-sequencing, these effects are minimized by sample-wide averaging and increased sequencing depth. However, single-cell data can pose a significant challenge during downstream analysis tasks because existing methods must consider the sparse nature of the data, and this might lead the existing methods to violate their assumptions.

Using zero-inflated versions of common distributions [30] is one way to address the issue of excessive zeros. It is still up for debate, though, whether scRNA-seq datasets generated using droplet-based techniques are zero inflated or standard distributions with lower means would better represent the extra zeros. Methods such as MAGIC [31], SAVER [32,33], and scImpute [34] can be used to impute parts of the zeros, replacing them with estimations of how truly expressed particular genes are based on their expression in similar samples. Nevertheless, imputation has the danger of creating a misleading structure that is not existent in the samples [35].

Typically, bulk RNA-seq studies contain preset sample groups, such as cancer cell lines and normal cell lines, various tissue types, or treatment and control groups. Similarly, it is feasible to arrange scRNA-seq research by classifying cells similar in nature into one group based on surface markers, sampling them at a succession of time periods, or comparing treatment groups. However, single-cell experiments are typically more investigative. Numerous single-cell investigations conducted to date have collected samples of developing or mature tissues and sought to profile the cell types present [36–42]. This strategy is best shown by the Human Cell Atlas project [43], which aims to compile a source for the transcriptional profiles of all the single cell lines present in humans. Similar studies have been conducted on specific tissues of other species [44].

A standardized analysis workflow applicable to numerous single cell experiments has been devised [45]. The workflow for the downstream analysis of single cell data is divided into several steps, namely, cell filtering, gene filtering and normalization.

1.2.2 Quality control of the libraries

The profiling of single cells during single cell RNA sequencing (scRNA-seq) is impacted by various technical and biological variations like cell specific capture efficiency, amplification bias, dropout events, stochastic gene expression, environmental niche, and cell cycle phases. The quality of a sequenced cell can be determined on the basis of two parameters. These parameters are library size and the number of expressed features in each library. Library size can be defined as the sum total of the expressions of all genes for a particular cell. If the library size of a particular cell is relatively small in comparison to other cells of the dataset then the cell is said to be of low quality. The reason behind the resultant low quality cell can be inefficient capturing of ribonucleic acid (RNA) molecule during library preparation. The other measure of cell quality, that is, the number of expressed features in each library can be defined as the total number of features with non-zero read counts for a cell. If the expression for most of the genes of a cell is 0 then that means the transcript remained uncaptured even though it was present during the sequencing process. Of note, the aforementioned parameters can be used to distinguish the bad quality cells from their good quality counterparts [45].

1.2.3 Filtering out low abundance genes

The single cell data is cursed with dimensionality both in terms of samples and features. Therefore removal of low abundance genes will minimize the computational requirement. Also, the low quality genes do not add any extra information to the information pool as their expression or count is zero or near zero for most of the cells due to dropout events. We can filter out low quality genes using two approaches, namely, mean based filtering and atleast *n* filtering. In the case of mean based filtering, we remove genes with a mean count below a particular threshold. Whereas in the case of atleast *n* filtering, we keep genes whose expression is non-zero in atleast *n* number of cells.

The mean based filtering is less rigid than atleast *n* filtering. If a gene is adequately expressed in some cells then it would not be filtered out. Also, mean based filtering helps in retaining genes that can be used in the identification of rare cell types. For example, if the gene *G* is highly expressed in some of the cells and has zero or near zero expression in the remaining cells then its average count will be above the filter threshold. Therefore, we will hold on to gene *G* for the identification of rare cells. On the other hand, in the case of atleast *n* filtering if the value of *n* is set to 10 and gene *G* is expressed in 9 cells then it will be filtered out regardless of its count in those 9 cells. Hence, atleast *n* filtering might lose outlier driven genes [45].

1.2.4 Normalization of cell specific biases

The raw single cell data can be normalized either by normalizing the genes or by normalizing the cells. Normalization of cells is performed to eliminate amplification biases introduced during sequencing protocols. Cell normalization can be accomplished via commonly employed read count normalization approaches, such as transcripts per million (TPM), fragments per kilobase million, and reads per kilobase million, which normalize each cell by the total amount of short reads plus a scaling factor. The biases which are related to amplification and sequencing depth are ignored by the unique molecular identifier (UMI)-based protocols as different reads linked with the particular UMI are condensed into a single count [46]. Normalization is useful for this sort of data [47] due to the fact that cell lines are typically not sequenced in saturation (i.e., an individual molecule is studied at least once). Alternatively, we can also perform cell normalization using spike-in sequences. The external RNA control consortium can be used as spike-in sequences based on the idea that technical factor affect the intrinsic and extrinsic genes equally [48]. It is also usual practise to perform log transformation after adding 1 as pseudo count to the read counts data [49].

Normalization of genes across samples is also undertaken to prevent highly expressed genes from dominating the study. For example, we can use z-score normalization [49] for normalizing the genes expression. Standardizing the features will help in the downstream analysis of the single cell data like cell type identification. After standardization, there will be an expression shift due to which genes will lose their relative scale and will be less zero inflated. This will lead to a change in the clustering performance of the single cell data.

A normalization component is provided by the SINCERA [50] pipeline for preprocessing scRNA-seq data. The package carries out cell normalization by trimmed mean and computes the z-score for gene normalization. One can use the visualization features, namely, MA plot, Q-Q plot, intersample correlation and distance measures of SINCERA pipeline to decide whether trimmed mean should be used to perform normalization or not.

On scRNA-seq data, some techniques conduct more specialised normalisation. For instance, during the clustering process, BISCUIT [51] employs iterative normalisation by learning parameters that describe the technical variations. Rare cell type identification (RaceID) [52] utilises median normalization to perform normalization. Instead of genes, transcript compatibility counts (TCC)-based clustering [47] uses equivalence classes as features and normalizes each feature by dividing it with the total read across all the cell lines.

The premise is that genes and cells with extremely low expression are considered to be erroneous signals in the data, so they are often removed from the library. Different thresholds have been created by previously published works for the removal of low quality samples and features and this might vary based on the number of cell lines and genes present in the data. For example, when single cell variance driven multitask clustering (scVDMC) [53] was applied for the analysis of droplet based peripheral blood mononuclear cell (PBMC) data, genes which are expressed in more than three cells and cells whose UMI count is equal to or more than 200 are considered for downstream analysis.

Although the majority of the existing clustering methods include global normalisation of genes and cells, the impact on the clustering findings is still up for debate. According to the analysis in [46], the use of bulk-based cell normalisation techniques can have major negative effects on the analysis of scRNA-seq data, such as the identification of highly variable genes prior to clustering in the face of high levels of biological noise and dropout events. If we use median of total read count of all cells or spike-ins to perform global normalization [51] then it might not solve the problem of dropouts rather eliminate biological stochasticity which is specific to a cell type which will eventually lead to biological conclusions which are not true in nature like improper clustering

1.2.5 Reducing feature space

Single cell RNA sequencing (scRNA-seq) allows us to profile thousands of genes at the resolution of a single cell. However, the large scale single cell datasets have their own share of advantages and disadvantages. If they offer an opportunity to study why the cells of the same tissue respond differently, they also possess challenges in terms of data analysis. The inadequacy of starting RNA causes quietening of genes during the polymerase chain reaction cycle which makes single cell data sparse with redundant features. Therefore, it is necessary to reduce the number of features that are redundant and less informative for efficient downstream analysis. There are various feature extraction and feature selection techniques present in literature that can be used to downsize the feature space to make it more informative and less computationally expensive.

Principal component analysis (PCA) linearly projects the high dimensional data into lower dimensional feature space. These low dimensional representations of data are independent and orthogonal to each other. In single cell consensus clustering (SC3), principal component analysis is applied to a cell-cell distance matrix to get eigenvectors corresponding to the top 15 eigenvalues for reducing the feature space [54]. The *pcaReduce* algorithm initially assumes that the data has more number of clusters *k* and then uses *k*-1 principal components to determine the *k* subsets. It iteratively reduces the number of principal components to determine the clusters in the data with *k* [55]. But, PCA cannot capture the non-linearity of the dataset.

The t-distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear dimension reduction technique. In the original feature space, tSNE models the Gaussian distribution of the neighboring samples around each sample. Similarly, in the reduced feature space tSNE models, the t-distribution of the neighboring samples around each sample such that the difference between the two distributions is minimized. FIt-SNE employs polynomial interpolation and fast Fourier transform to compute low dimensional embedding of the single cell data without downsampling [56]. However, one of the limitations of t-SNE is that it requires parameter tuning.

Recently, various deep learning based approaches which employ neural networks and autoencoders to interpret high dimension single cell data have been proposed. One such method is SCNN [57] which uses neural network architecture infused with biological knowledge to get the low dimensional representation of the single cell data. Another method SCVIS [58] is based on a variational autoencoder and learns a latent variable model to reduce the feature space while preserving the distance between cells in the low dimensional space.

The issue with the feature extraction techniques is that they do not allow interpretability by a domain expert whereas feature selection techniques provide important gene markers which can be further studied by a domain expert to draw important inferences. Some of the popular feature selection techniques used in literature are coefficient of variation [59], Gini index [60], and PCA loadings [61]. The coefficient of variation selects features depicting high variability among cells for further analysis. Gini index is computed using the Lorenz curve which is a visual representation of statistical dispersion. Features with high Gini index value are selected during the feature selection. In the case of PCA loadings, the Pearson correlation between the features and principal components is computed. The features are sorted based on the correlation value and the features with high correlation are selected for further analysis.

1.2.6 Cell characterization

The popular single cell clustering techniques used in literature to determine the hidden structure of the single cell data are Seurat [62], GiniClust [60], and single-cell consensus clustering (SC3) [54].

Seurat uses the ratio of the variance of a gene's expression to its mean for determining highly variable genes. The selected genes are then used to compute a distance matrix of cells or samples. This distance matrix is used to construct the shared-nearest neighbor graph (SNN) which is further passed as an input to the smart local moving (SLM) algorithm for clustering.

GiniClust employs the Gini coefficient to select highly variable genes. Gini coefficient is a mathematical representation of the statistical dispersion which is computed using a Lorenz curve. In the Lorenz curve, the cumulative percentage of cells is represented on the x-axis, and the cumulative percentage of a gene's expression is represented on the y-axis. In the idle case, if each cell has the same amount of mRNA expression corresponding to a gene then we will get an upward sloping linear line which represents the perfect equality, i.e., 1. To compute the Gini coefficient for each gene, a ratio of the area between the line of equality and Lorenz curve to the total area beneath the line of perfect equality is taken into consideration. Genes with high Gini index values are selected. A density-based clustering algorithm is employed on the selected genes to determine the clusters in a dataset.

SC3 uses a cluster-based similarity partitioning algorithm (CSPA) to identify the cluster of cells in the dataset. In SC3, principal component analysis is applied to a cell-cell distance matrix to get eigenvectors corresponding to the top 15 eigenvalues. K-means clustering is carried out on the eigenvectors. A similarity matrix is computed based on the output of the K-means algorithm. If two cells belong to the same cluster then their similarity will be 1, otherwise 0. In the end, an average of the similarity matrices generated from each iteration of K-means is used to form the consensus matrix. Hierarchical clustering is applied to the consensus matrix to determine the similar type of cells in the dataset.

1.3 Low dimensional representation of biological entities

The count of manuscripts related to biomedical research which are published daily in peer-reviewed journals is around 3000 [63]. It is very strenuous for researchers around the world to keep track of the findings published in these manuscripts. This resulted in the need for low dimensional representation of biological entities (like genes, diseases and tissues) in the form of word vectors. When word representation models are used to create these word embeddings, they can eventually be used to pull information from the huge corpora to find out how the pathological activities of tissues change when a gene's normal activity changes.

1.3.1 Word representation models

The Word2vec model of word vectors was created by Mikolov and colleagues [64]. A shallow neural network with two hidden layers is used by this model to make a vector for each word. The Continuous Bag of Words (CBOW) and skip-gram versions of word2vec are meant to capture the cognitive and linguistic information of words in the word vectors. Word2Vec uses one of the two neural network variants, that is, CBOW and skip-gram, to make predictions according to the context. In both methods, a preset length of the frame is shifted in tandem with the corpora, and learning is performed with words within the frame at each iteration [65]. An effective tool for revealing associations in the text and measuring token similarity is provided by this feature presentation method. This approach, for example, would place the terms "small" and "smaller" close to one another in the feature space.

Continuous Bag of words (CBOW) provides context-based word recommendations for present work. CBOW connects with the surrounding window terms. The CBOW technique makes use of three levels. The first or input layer is context, and the hidden layer corresponds to the approximation of each word from the first layer to the weight matrix, which is then approximated to the final, which is the third and output layer. Based on the backward propagation of the error gradient, the final phase of this method correlates the result with the task itself in order to improve the approximation. While the skip-gram approach bases its prediction of the context word on the center word, the CBOW method bases its prediction of the center word on its context [66].

The Skip-Gram approach is the opposite of the CBOW approach; following context training, predictions are made based on the center word. The targeted word is correlated in the input layer, and the context is correlated in the output layer. Unlike CBOW, this approach seeks to determine the context of a word. This model's final stage is the association across the output and each word in the context to modify the depiction via backward propagation. When training data is sparse and uncommon terms are well-presented, skip-gram is effective. CBOW, in contrast, is faster and provides better performance with repeated terms [67].

Bidirectional Encoder Representations from Transformers (BERT) [68] is a different language paradigm for contextualized word vectors in which concurrent attention layers are used by the transformer neural network instead of linear recurrence. BERT is learned with two activities in place of the fundamental learning activity to promote bidirectional prediction and sentence-level comprehension.

Prior to BERT, a frequent language modeling goal was to anticipate the following word (given a series of words) [69]. The cloze exam served as inspiration for BERT's additional language modeling goal, which was utilized to build the model [70]. This goal calls for the model to anticipate a "masked" token rather than anticipate the following token, thus the term Masked Language Modeling (MLM). MLM arbitrarily replaces 15% of the input sequence's tokens with special tokens, such as "[MASK]," to mask out some of the terms. For instance, the sequence “the single cell data is sparse in nature” is changed to “the [MASK] cell data is sparse [MASK] nature”. Based on the knowledge present in the sentence's unmasked tokens, the objective would be to forecast the masked (absent) tokens. This enables the model to include conditioning that takes into account information from both the right (forecasting the future token) and left (forecasting the previous token) sides of the token. In order to be more exact, BERT uses a more thorough masking method since the [MASK] token only occurs at the time of training the model. BERT occasionally substitutes the term with an arbitrary term or the exact term which has an 10% chance, as opposed to always substituting it with the special [MASK] token which has an 80% chance. It's significant to notice that the algorithm is not provided details about absent words (or words that have been substituted by different word). The percentage of this modification (for example, 10% of the total input size) is the only piece of information that is provided to the algorithm. The model must predict these terms and offer suggestions for replacements. The goal allowed BERT to overcome the unidirectional restriction of previous models and incorporate information from both left and right side of the token.

A Next Sentence Prediction (NSP) goal is used in BERT in addition to the MLM goal. In this task, the model must determine whether or not a particular sentence can be regarded as the following phrase to the present sentence. This is driven by the requirement for the framework to integrate links between phrases or to use knowledge outside the bounds of a phrase in order to perform well in specific tasks. It's a binary classification problem. The model is given a following phrase *Y* for each phrase *X*, and is then questioned if *Y* follows *X*? The real following sentence is swapped out with an arbitrary phrase 50% of the time to provide a consistent self-training set of data. This goal aids the algorithm in understanding the connections among phrases and has been demonstrated to be helpful in applications like question answering and natural language interpretation [68].

BERT's training goal is to reduce the overall training error of MLM and NSP. Learning for BERT is done using pairs of sequences (provided the NSP objective). BERT utilises two unique tokens, [CLS] and [SEP], to differentiate among the two input sequences. The [CLS] symbol is added at the start, and the [SEP] symbol is added in the room between the two lines. The encoder then receives the complete sequence. The [CLS] token's outcome, utilized in a softmax function for the classification, stores information about the NSP result. The initial BERT is learned in Base and Large settings. The number of attention heads, size of the representation, and number of encoder layers vary between the two versions.

1.3.2 Biomedical hypothesis generation

The purpose of making predictions is to uncover hidden associations that are implied by the existence of other apparent associations but are not explicitly stated in the literature. Specifically, the process of relating two previously unrelated pieces of knowledge is referred to as hypothesis creation. For instance, it might be known that chemical B, which causes sickness A, is reduced by medicine C, which is known to do the same. However, the connection between sickness A and drug C may not be known because the corresponding studies were authored independently of one another (a situation known as “disjoint data”). The goal of hypothesis generation is to find these hidden relationships in biomedical literature.

Text mining on biomedical literature has produced a number of breakthroughs. The tool BioWordVec finds similar sentences by integrating the subword information from unlabelled biological content with Medical Subject Headings (MeSH) [71]. In the work, the authors have used the low dimensional word representations to determine whether a person can suffer from a mental illness in the future based on their responses on the social media platforms. The language model was also used to find meaningful connections between different brain regions by *Rosenthal et al.* [72]. Word embeddings have also been used in analyzing and identifying novel medicinal advantages of supplements [73], as well as prospective illness treatments [74]. There are numerous examples present in the literature where word representation models were successfully employed in the quest of finding solutions for various biomedical problems like compound-protein interactions [75] and drug discovery [76].

Chapter 2

SelfE for identifying significant features in single cell data

2.1 Introduction

Recent technological advances in the field of genomics have made it possible to sequence messenger RNA molecules from each cell line. Single-cell RNA sequencing (scRNA-seq) provides insight into the variability of a tissue's transcriptome, which is otherwise averaged out in bulk RNA-seq experiments [77]. For instance, a tumour may have tens of clones that exhibit substantial phenotypic variety [38] [78] [79]. The consequences of averaging prevent the bulk expression profile from capturing this variance. Recent advances in single cell transcriptomics have uncovered baffling intra- and inter-tumor genomic variation in numerous cancer types [38] [78] [80]. Despite the undeniable utility of single-cell transcriptomic studies, data retrieved from numerous high-throughput single-cell platforms are subject to large dropout rates due to the scarcity of starting RNA. This makes the study of scRNA-seq datasets unconventional and difficult. Additional complications are introduced due to trivial biological noise and intrinsic technical noise in the sequencing phase. Single-cell transcriptomics necessitates fresh statistical and computational approaches to surmount the new problems posed by its considerable distinctions from bulk transcriptomics.

Enhancing the quality of expression-based clusters of individual transcriptomes has been greatly facilitated by feature engineering techniques. It is typically grouped as feature extraction and feature selection strategies. Techniques for feature extraction modify the feature space. Feature selection strategies, on the other hand, reduce feature space by selecting meaningful and non-redundant features. Several well-known feature extraction methods for the examination of single-cell RNA-sequencing data are as follows: Principal Component Analysis (PCA) [30], t-distributed stochastic neighbor embedding (t-SNE) [81], Zero-Inflated Factor Analysis (ZIFA) [30], Zero-Inflated Negative Binomial-based Wanted Variation Extraction (ZINB-WaVE) [82], and scvis [58]. The first two principal components of PCA are proportional to the number of genes found per cell. However, it is possible that a certain gene is expressed, but not discovered, due to many rounds of amplification during sequencing of the input substance. Because of its high time and spatial complexity, t-SNE becomes less scalable when applied to large datasets, despite its widespread application in single-cell research. Although ZIFA takes dropout occurrences into account, it does not examine the count nature of data [82]. In addition, ZIFA estimates the relevance of genes based on their average expression level which deteriorate with an exponential rate. ZINB-WaVE, on the other hand, takes into consideration both the dropout events and count nature of the data, however the issue with feature extraction approaches is that it is challenging to infer any physical meaning for interpretation from the changed space. Therefore, this work gives a technique for selecting features as it explains the concept of smaller feature space.

On the basis of evaluation criteria, feature selection strategies are further classified as follows: filter [83], wrapper [84], and embedded techniques [85]. In case of filter methods, the importance of each feature is computed, and the feature(s) with importance below a threshold value are disregarded. In contrast to wrapper approaches, in which each feature subset generated from the space of possible feature sets is evaluated by training and testing a classification model, embedded techniques incorporate the search for the optimal feature subset within the model development [86]. Although wrapper approaches take into account the performance of the learning algorithm for feature evaluation, they are computationally expensive and have a high risk of overfitting because the classifier is repeatedly called for the evaluation of the feature subsets. Similar to wrapper techniques, embedded techniques also depend on the modeling algorithms. As unsupervised filter-based gene selection approaches are independent of annotations and modeling algorithms, they will be the focus of this research.

Some current state-of-the-art approaches that have been considered as benchmarks include the coefficient of variation [59], Fano factor [26], Gini index [60], and PCA loadings [61]. In coefficient of variation, bins are made, and then the mean value of each feature present in the high dimensional space is computed. The features are then allotted to a bin based on its mean expression level. The ratio of median to the median absolute deviation of the dispersion index of each feature present in a bin is calculated. The bin's features with the highest ratio of median to median absolute deviation are considered for further analysis and are considered to be more variable in nature in nature in comparison to the other bin’s features. We, therefore, choose features with high levels of normalized dispersion [59].

The Lorenz curve and Gini coefficient are metrics of statistical dispersion. The Gini index is a quantitative indicator that is calculated using a Lorenz curve, a visual indicator of disparity. The x-axis of the Lorenz curve for feature selection shows the cumulative percentage of the cell or sample population, and the y-axis shows the cumulative percentage of gene expression for a specific gene. When plotting the percentage of transcripts over the percentage of cells, the upwardly sloping straight line reflects perfect equality. Consequently, in the Lorenz curve, the actual distribution of gene expression is compared to the line of perfect equality. An equal distribution of gene expression is indicated by a smaller distance between the Lorenz curve and the line of perfect equality. The mathematical indicator known as the Gini coefficient converts this visual interpretation into a scale where 0 denotes perfect equality, and 1 denotes perfect inequality. Gini index is calculated as a ratio of the entire region below the line of perfect equality to the region between the line of equality and the Lorenz curve. Higher Gini index genes are chosen.

The Pearson correlation between the feature and components is taken into account via a different method called PCA loadings for feature selection. Among the first *n* component scores, the highest score is assigned to the features. The features are then sorted in descending order based on these component scores. During feature selection, the top features are chosen. However, the drawback of PCA loadings is that they presuppose that the data has a gaussian distribution. But, dropout occurrences cause the single-cell RNA sequencing data to be non-gaussian [30].

The Fano factor [26], which is characterized in the literature as the ratio of variance to the mean, is another prominent technique for feature selection. The Fano factor of genes will vary less in an unbalanced population where a given sample type makes up less than 0.01% of the total population.

All of the aforementioned methods [26,59,60,87] are largely founded on the idea of heterogeneity and variability, i.e., genes should be chosen based on how well they can account for the variation in the data. Up until now, this has been the widely held belief. In this study, we put out an unique supposition. According to our theory, genes ought to be chosen based on their potential to 'express' other genes. Assuming that genes are interdependent, the majority of genes can be expressed using a limited number of genes. The goal of this research is to choose a small number of genes that can convey all other genes.

The genes are assumed to be linearly dependent on one another in order to model the expressibility mathematically. In this work, this is referred to as "self expression." We use a sparse selection approach to pick the genes that best account for the expressibility based on the linear dependence model. We acknowledge that the linear dependency hypothesis might be oversimplified. The dependency may not always be linear.

2.2 Methods

2.2.1 Proposed algorithm

Our technique for gene selection is predicated on the premise that gene expressions are linearly dependent, i.e.,

$$x_i = X_{i^c} c_i \forall_i \tag{2.1}$$

In the above equation, the vector x_i describes the i^{th} gene of the single cell data and the matrix X_i^c represents the remaining genes present in the single cell data. The weights c_i represent the gene x_i as a linear combination of the remaining genes. This applies to all genes.

The main idea is to select genes that can illustrate other genes as linear combinations of the selected genes. The weight vector, that is, $c_{i,}$ must have a structure in order to determine the gene subset which can represent other genes of the feature space. We can explain this better using equation 2.1 when it is represented in the form of genes.

$$X = XC,\ such\ that\ C_{ii} = 0 \tag{2.2}$$

In *X*, all the genes of the scRNA-seq data are arranged in columns whereas in *C* all the c_i's are arranged as columns with 0 as the diagonal element. To keep the gene from expressing by itself, this restriction is required.

Now, it will be clear how *C* is structured as mentioned. To select a subset of genes that can explain other genes of the feature subset, *C* needs to have a row sparse structure. In equation 2.2, we need to express the genes of *X* on the left hand side using a few genes from the *X* on the right hand side, which means the rows of *C* corresponding to the selected genes must be non-zero other than the diagonal element. But, as we need only a handful of genes to represent the entire feature set, most of the rows of *C* need to be zeros which makes the structure of *C* a row-sparse structure.

This is called a row-sparse multiple measurement vector recovery problems in mathematics. This is illustrated as

$$arg\ min_c\ ||C||_{2,0}\ such\ that\ X\ =\ XC \tag{2.3}$$

The $l_{2,0}$-norm refers to the number of rows in *X* that are not zero. In order to ensure that the *X* on the right hand side has no zero value, we apply the l_2-norm on the *X*. It is necessary to ensure that *X* on the right hand has non-zero values because we need all the genes on the left hand side of equation 2.2 to be expressed. The l_0-norm on *X* ensures that only a few rows are selected from the *X* on the right hand side, as we need only a handful of genes to illustrate the entire feature space. It is well known that the problem described in equation 2.3 is NP (non-deterministic polynomial time) hard. However, it can be roughly resolved by a greedy method known as simultaneous orthogonal matching pursuit (SOMP) [88].

SelfE Algorithm

```
procedure SelfE ( data, k )
   Initialize: idx = [ ], Y = data, Phi = data, R = Y
   For loop, iterate ( Number of features )
      K = abs ( R' * R )
      For loop, iterate ( Number of features )
         c ( i ) = norm ( K ( i, : ) )
      End loop 2
      [ val, pos ] = sort c in descending order
      idx = [ idx  pos ( 1 ) ]
      Extracted feature is multiplied with its pseudo inverse and original data to minimize error
   End loop 1
```

2.2.2 SOMP algorithm

The SOMP algorithm for solving the row-sparse recovery problem is well known in the signal processing community. However, for the benefit of the reader, we discuss its mechanism intuitively. It is used for solving problems of the form -

In the field of signal processing, the SOMP algorithm is well recognized for resolving the row-sparse recovery issue. However, we outline its functioning simply for the reader's advantage. It is used to accomplish tasks of the following form:

$$Y = AX \tag{2.4}$$

In the above equation 2.4, Y and A are given, whereas X is to be estimated. Additionally, the row-sparse structure of solution X is well known.

Since the structure must be preserved, conventional methods like the Moore-Penrose pseudoinverse do not produce the desired result. Therefore, we need to use particular algorithms like SOMP. The fundamental principle of SOMP is to identify the support gradually, that is, the position of non-zero rows of the solution X and estimate the values that go with it.

In certain circumstances (which we omit for the sake of the reader's general interest), SOMP presumes that A behaves as an approximate isometry; that is, if we multiply transpose of A with A then we will get an identity matrix. Keeping this assumption in mind, we shall describe the SOMP mechanism with a toy example.

Consider a scenario in which X comprises 16 rows, of which only rows 2 and 9 contain non-zero values while the remaining contain zeros. As a result, locations 2 and 9 will have peaks when we plot the l_2-norm of each row. If we pre-multiply equation 2.4 with $A^{T,}$ then we will get the following equation:

$$A^T Y = A^T AX \tag{2.5}$$

As was previously noted, A serves as an approximate identity, therefore we will obtain high values at rows 2 and 9 when l_2-norm is applied , and zeroes everywhere else. This is the initial phase of SOMP. SOMP chooses the position with the greatest magnitude, in this case, it is position 2. Once the position has been identified, the value at the particular location must be determined. This is accomplished by minimizing the l_2-norm as follows:

$$X_\Omega \leftarrow min_X \left|\left|Y - A_\Omega X\right|\right|_F^2 \tag{2.6}$$

According to this formula, SOMP determines the values at the points chosen in the support set Ω. This concludes one cycle of SOMP.

The first non-zero repeat's location was already discovered in the first iteration. The same location shouldn't be repeated in later iterations. Consequently, the answer must be orthogonalized. This is easily accomplished using -

$$R = Y - A_\Omega X_\Omega \tag{2.7}$$

We pre-multiply A^T with R in place of Y in the second iteration to avoid selecting indices that are already available in the set Ω. l_2-norm is applied to the rows, and the rows which result in the highest value are selected and included in the support set Ω. Finding the values at the places in the support set is the next step. To do this, we need to solve equation 2.6.

Since there are only two non-zero places in our hypothetical example, the method terminates after two iterations. However, this discussion is sufficiently general for the reader to comprehend what transpires when the size of the support set increases.

2.2.3 Kernel SelfE

So far we have discussed the linear version of our algorithm. But what if, the samples cannot be expressed as linear combination of other data points? That is, what if the relationship is not linear? The standard approach in machine learning to handle such issues is to employ kernelization. The samples are non-linearly projected onto a higher dimension (possibly infinite dimension) and it is assumed that in the projected space, the samples can be expressed as a linear combination of other datapoints/samples. In practice, it is not possible to know the higher dimensional non-linear projection, so the problem is solved by the so called kernel trick; instead of trying to figure out the projections, the inner products are kernelized to emulate the effect of nonlinearity. The approach will be clearer when we elucidate it with our problem.

The basic model has been expressed in equation 2.2

$$X = XC,\ such\ that\ C_{ii} = 0$$

If we assume that the data has beeen projected non-linearly to a higher dimension, we can express equation 2.2 as follows

$$\psi(X) = \psi(X)C,\ such\ that\ C_{ii} = 0 \qquad (2.8)$$

where ψ is the non-linear projection

This can be equivalently expressed as follows, after pre-multiplication by $\psi(X)$

$$\psi(X)^T\psi(X) = \psi(X)^T\psi(X)C, \text{ such that } C_{ii} = 0 \qquad (2.9)$$

The equations 2.8 and 2.9 are equivalent, i.e., the solution C is equivalent for both cases. But now equation 2.9, is directly amenable to the kernel trick. The kernel is defined on the inner product,

$$K = < \psi(X), \psi(X) > \qquad (2.10)$$

Therefore, we can express equation 9 as

$$K = KC, \text{ such that } C_{ii} = 0 \qquad (2.11)$$

This is the kernelized version of SelfE. The formulation equation 2.11 is the kernelized version of equation 2.8. The only difference between the two is that instead of working on the raw data matrix *X*, we are using the kernelized version of the data matrix *K*.

The structure of C remains the same as before. I.e., it will have a row-sparse structure. Therefore, one can solve it using $l_{2,0}$-minimization.

$$min_C ||C||_{2,0} \text{ such that } K = KC \qquad (2.12)$$

As discussed before, this can be solved by SOMP.

2.3 Datasets used for validation

2.3.1 Dataset description

In this study, four distinct single-cell RNA-sequencing (scRNA-seq) datasets were utilized to compare the performance of SelfE to that of conventional gene selection techniques [26,59–61].

- **Cell line:** Fluidigm, a procedure based on microfluidic technology, was utilized to perform scRNA-seq on 630 single cells extracted from seven cell lines. As separate sequencing was done for every cell line, therefore original annotations were linked topically. Nine distinct cell lines are produced by the sequencing - A549, GM12878 B1, GM12878 B2, H1 B1, H1 B2, H1437, HCT116, IMR90, and K562. Two distinct batches of the cell lines GM12878 and H1 were generated [78].

- **Jurkat - 293T:** This particular collection includes 3,300 transcriptomes taken from two distinct cell lines, namely, Jurkat and 293T cells. The transcriptomes are mixed in an artificial environment in quantities that are equal to one another (50:50). The variations and levels of the cell-type-specific identifiers CD3D and XIST are used to label all transcriptomes [89].

- **PBMC:** This dataset is comprised of 68,000 peripheral blood mononuclear cell (PBMC) cell lines from healthy volunteers. According to correlation with purified bulk RNA-Seq data derived from fluorescence activated cell sorting (FACS) of typical PBMC subtypes, they are categorized into 11 typical PBMC subtypes. For this investigation, we randomly chose 100 cells from every identified subtype and preserved the whole cluster if the cell count in the particular cluster was under 100 [89].

- **Usoskin:** 799 cell lines from mouse lumbar dorsal root ganglion (DRG) make up this dataset. The cells were grouped by the authors using an unsupervised method. 622 cells out of 799 were categorised as neurons, 68 cells had an uncertain classification, and 109 cells were annotated as non-neuronal. On the basis of prominent cell biomarkers, the 622 mouse neuron cells were further categorised into four primary groups: neurofilament containing (NF), non-peptidergic nociceptors (NP), peptidergic nociceptors (PEP), and tyrosine hydroxylase containing (TH) [40].

2.3.2 Steps followed for data preprocessing

The pipeline followed for preprocessing single cell data is depicted using Figure 2.1.

- **Data filtering:** As there were a few genes whose transcripts were not found in Cell Line and Usoskin, cell screening was not needed for the Cell Line and Usoskin. Low total read count cells are removed from Jurkat and PBMC samples. Considered filtered genes are those with a read count greater than 2 in at least 4 cells. The remaining genes are removed if they don't meet the aforementioned requirements.

- **Median normalization:** Each cell's median total read count is calculated. The ratio of a cell's total read count to the median of all read counts is used to calculate the expression of genes associated with that cell.

- **Log transformation:** The median normalized count matrix is transformed after 1 is added as a pseudocount.

- **Gene selection:** A log-transformed matrix is provided as input to SelfE in order to execute gene selection, along with the number of features to be chosen.

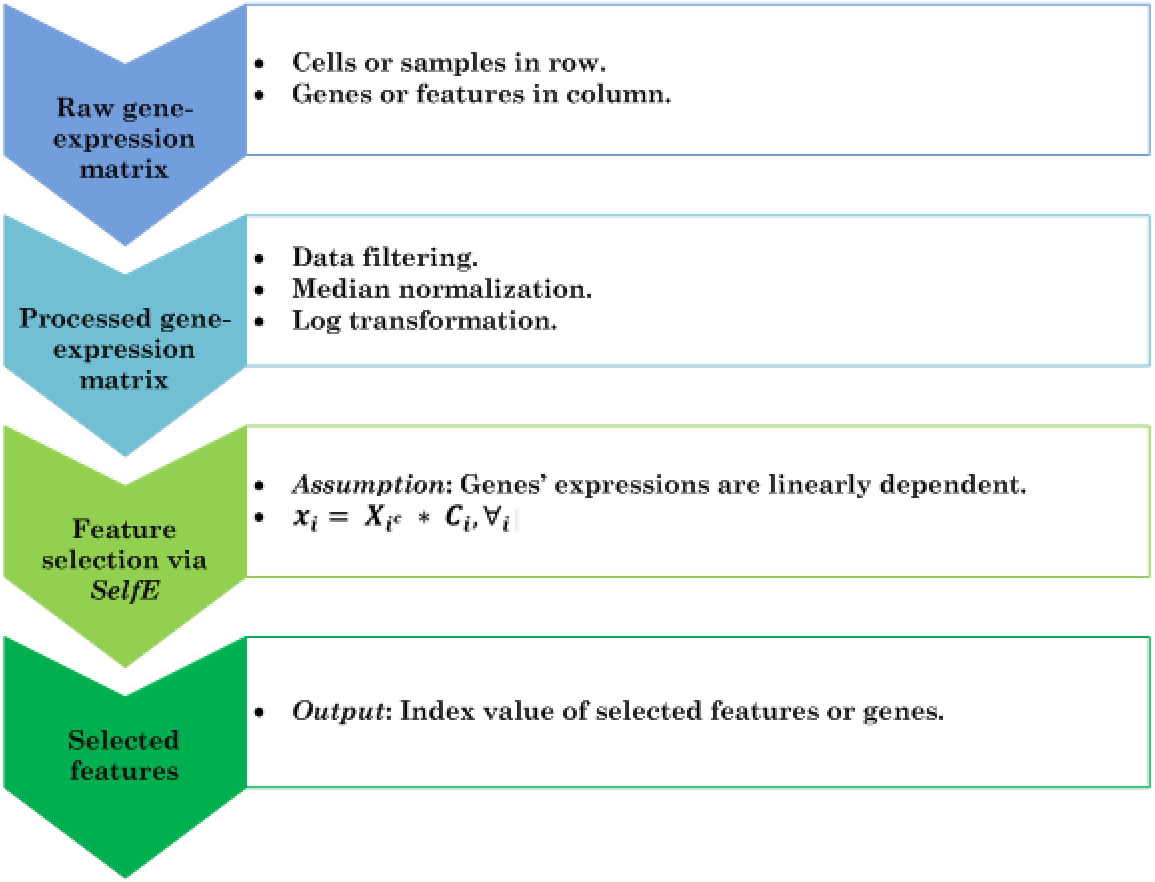
Raw gene-expression matrix
Cells or samples in row.
Genes or features in column.
Processed gene-expression matrix
Data filtering.
Median normalization.
Log transformation.
Feature selection via *SelfE*
Assumption: Genes' expressions are linearly dependent.
$x_i = X_{i^c} * C_i, \forall_i$
Selected features
Output: Index value of selected features or genes.

Figure 2.1: Breakdown of the SelfE Pipeline. Method and Experiments provide additional information.

2.4 Results

This section compares the suggested gene selection algorithm to other approaches in order to show how effective it is when compared to them. Comparative gene selection strategies were based on the coefficient of variation, the Fano factor, the Gini index, and the PCA loadings.

2.4.1 Validation based on clustering accuracy

The ideal number of genes to use for the categorization of single-cell data is a challenging question. We evaluated the effectiveness of clustering on both less and more extensive collections of genes for each set of data. We examined smaller gene sets with 20, 30, 40, and 50 genes as well as bigger sets with 100 to 500 genes spaced by 50 genes. With the use of SelfE, coefficient of variation, Fano factor, Gini index, and PCA loadings, the genes were chosen from each of the four datasets.

We used the Single Cell Consensus Cluster (SC3) to assess the selected genes acquired utilizing various gene selection approaches [54]. The SC3 clusters and the cell annotations were compared. For the same, we supplied the Normalized Mutual Information (*NMI*) [90] and the Adjusted Rand Index (*ARI*) [91]. In order to assess the clustering accuracy, *ARI* and *NMI* essentially calculate the ratio of the total number of sample pairs that are members of the identical cluster/different cluster in the actual and forecasted partitions to the total number of sample pairs. Each of these measures has values ranging from 0 to 1. When the original and anticipated clusters agree exactly, the ARI/NMI result is 1.

SelfE's performance was shown to be the most reliable across datasets. Notably, it outperformed the other approaches quite favorably, especially for the smaller number of genes. But as the gene set kept growing, we noticed that its performance flattened down. Noticeably, its overall accuracy was significantly superior on cell line datasets (called Cell Line and Jurkat data, respectively), where cell labels were most accurate (Dataset Description) (Figures 2.2 - 2.9).

The laplacian kernel was used to imitate the effect of non-linearity by kernelizing the inner products of samples. But when Kmeans clustering was employed to evaluate the selcted genes acquired using SelfE and Kernel SelfE, no appreciable improvement in the clustering measures of Kernel SelfE was observed (Table 2.1).

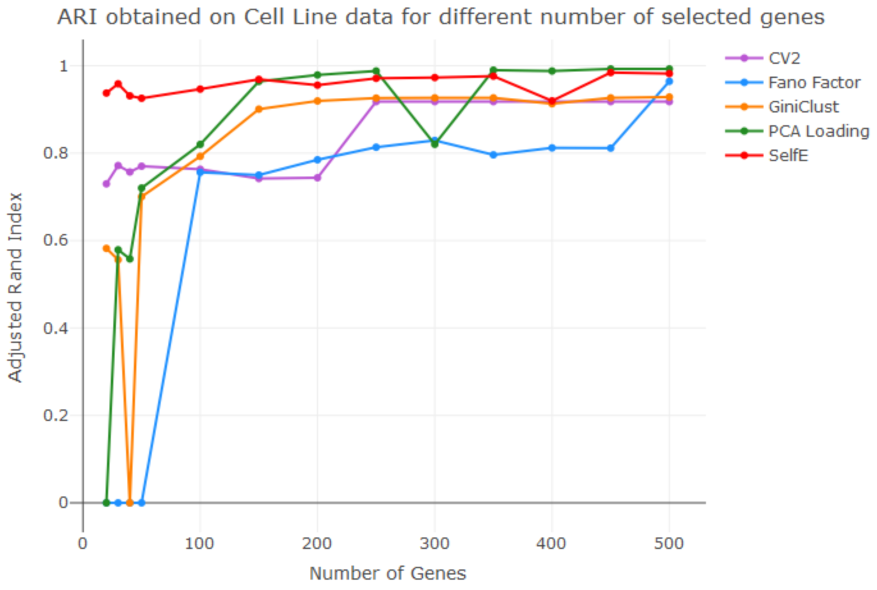

Figure 2.2: Each line of the line plot indicates a distinct approach used to choose a variable number of attributes. The x-axis represents the number of genes included in the feature subset, while the y-axis represents the associated adjusted rand index (ARI) determined using the SC3 algorithm in the case of cell line data.

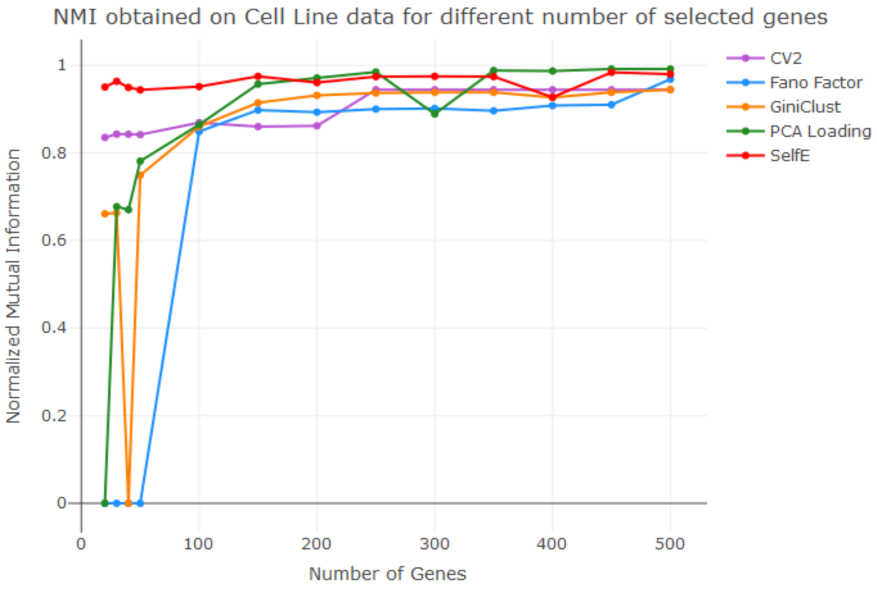

Figure 2.3: Each line of the line plot indicates a distinct approach used to choose a variable number of attributes. The x-axis represents the number of genes included in the feature subset, while the y-axis represents the associated normalized mutual information determined using the SC3 algorithm in the case of cell line data.

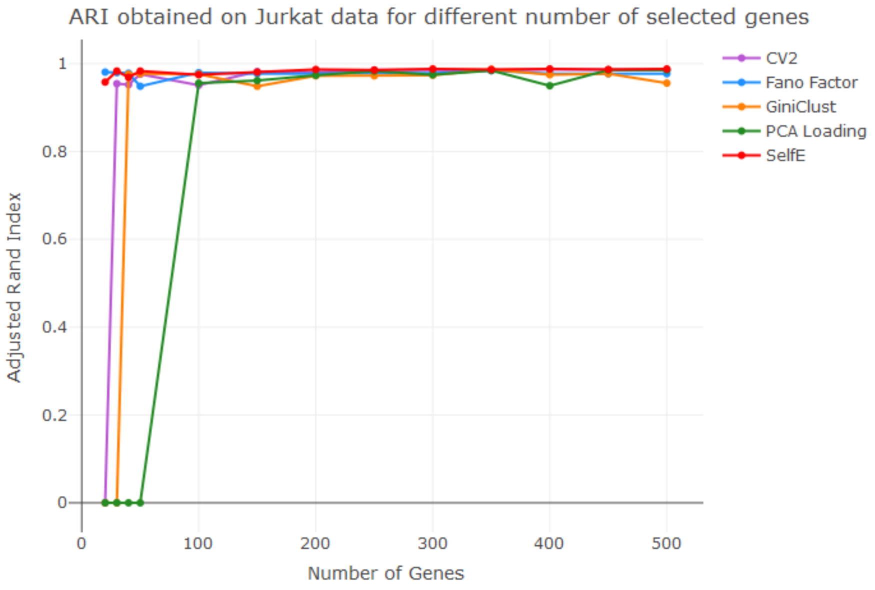

Figure 2.4: Each line of the line plot indicates a distinct approach used to choose a variable number of attributes. The x-axis represents the number of genes included in the feature subset, while the y-axis represents the associated adjusted rand index (ARI) determined using the SC3 algorithm in the case of jurkat - 293T data.

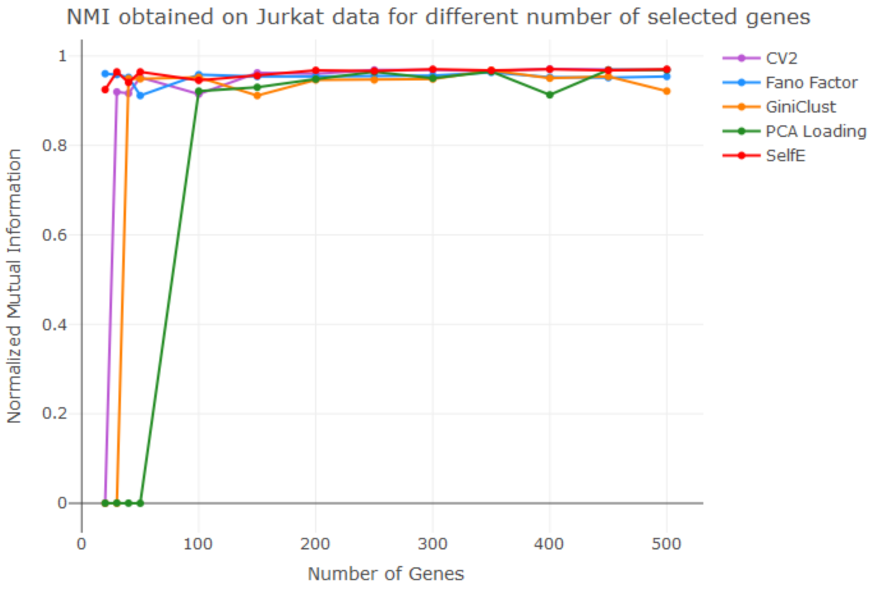

Figure 2.5: Each line of the line plot indicates a distinct approach used to choose a variable number of attributes. The x-axis represents the number of genes included in the feature subset, while the y-axis represents the associated normalized mutual information determined using the SC3 algorithm in the case of jurkat - 293T data.

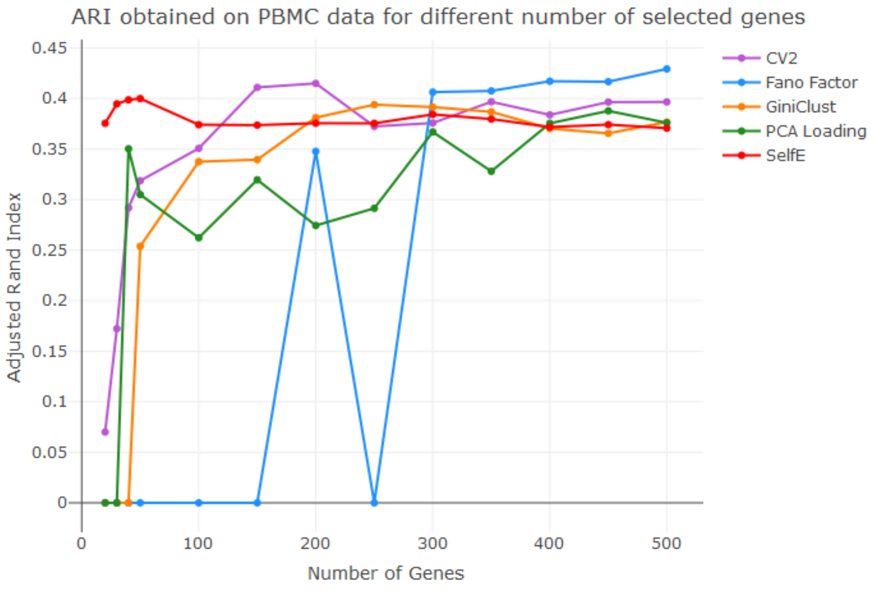

Figure 2.6: Each line of the line plot indicates a distinct approach used to choose a variable number of attributes. The x-axis represents the number of genes included in the feature subset, while the y-axis represents the associated adjusted rand index (ARI) determined using the SC3 algorithm in the case of PBMC data.

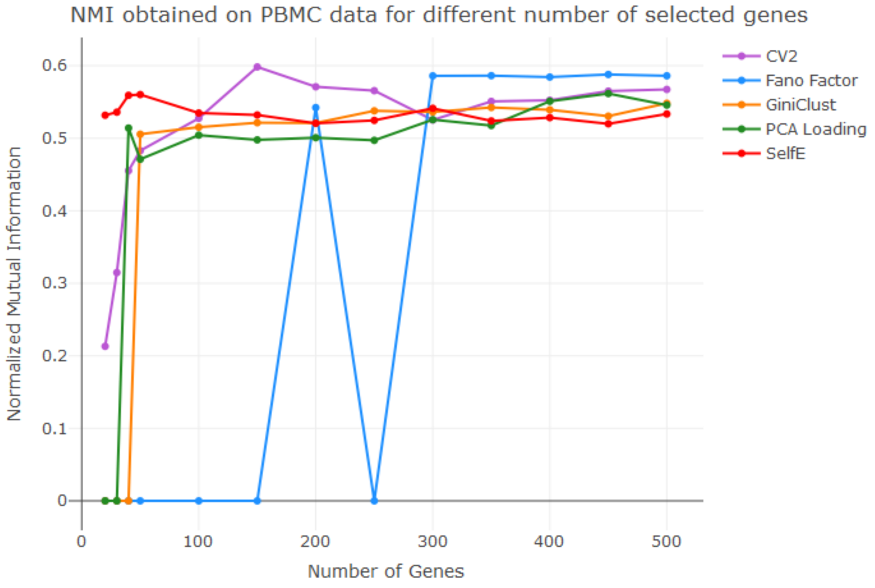

Figure 2.7: Each line of the line plot indicates a distinct approach used to choose a variable number of attributes. The x-axis represents the number of genes included in the feature subset, while the y-axis represents the associated normalized mutual information determined using the SC3 algorithm in the case of PBMC data.

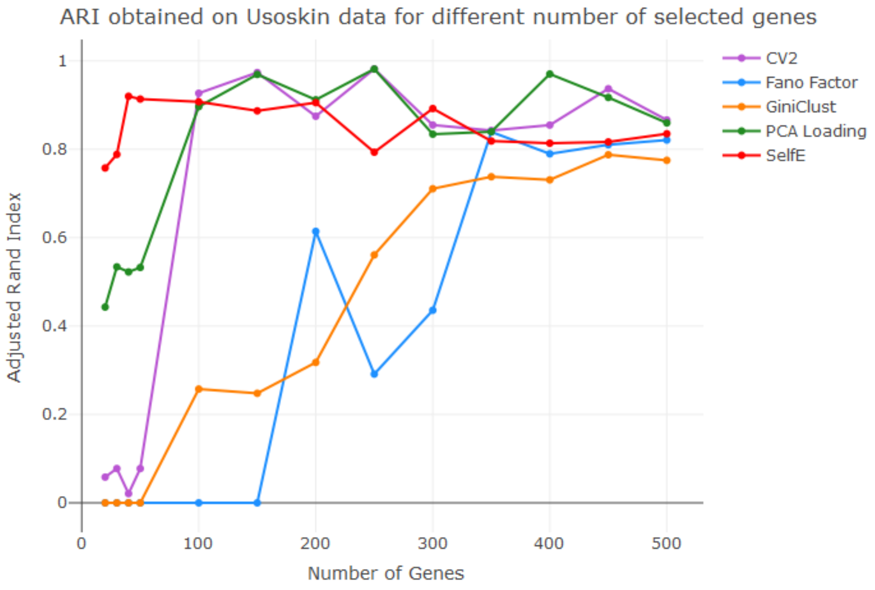

Figure 2.8: Each line of the line plot indicates a distinct approach used to choose a variable number of attributes. The x-axis represents the number of genes included in the feature subset, while the y-axis represents the associated adjusted rand index (ARI) determined using the SC3 algorithm in the case of usoskin data.

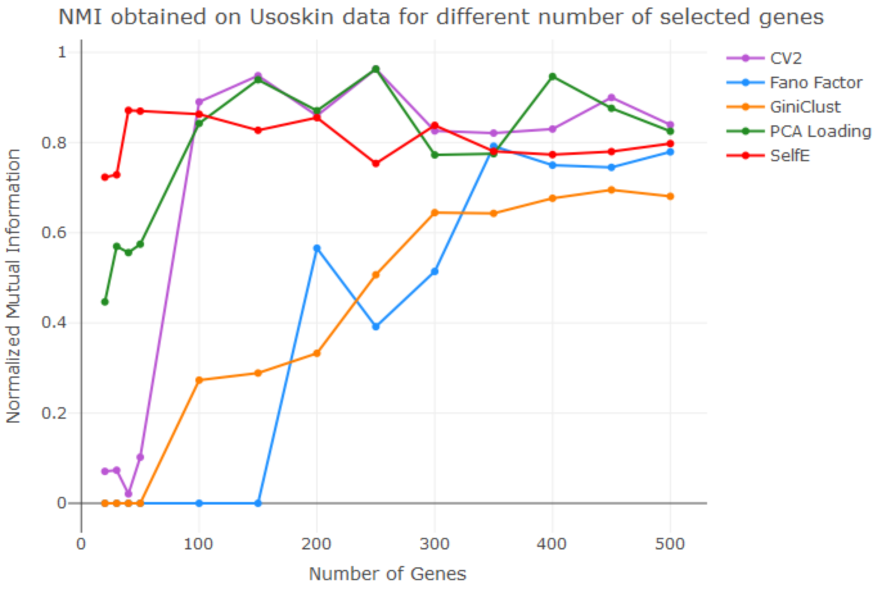

Figure 2.9: Each line of the line plot indicates a distinct approach used to choose a variable number of attributes. The x-axis represents the number of genes included in the feature subset, while the y-axis represents the associated normalized mutual information determined using the SC3 algorithm in the case of usoskin data.

Table 2.1: Comparison of clustering accuracy measures, Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), obtained on the different data sets using SelfE and Kernel SelfE for different gene sets

Feature Selection Technique		Datasets							
		Cell Line		Jurkat		PBMC		Usoskin	
		ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI
SelfE	20	**0.8276**	**0.8874**	**0.9819**	**0.9625**	**0.3840**	**0.5530**	**0.7421**	**0.7018**
	30	**0.8040**	**0.8658**	**0.9843**	**0.9667**	**0.3846**	**0.5663**	**0.7348**	**0.7015**
	40	0.7745	0.8464	**0.9806**	0.9588	**0.3875**	**0.5724**	**0.7497**	**0.7113**
	50	0.7839	0.8463	0.9794	0.9557	0.3788	0.5611	**0.6929**	**0.6806**
Kernel SelfE	20	0.7731	0.8678	0.9770	0.9497	0.3298	0.5294	0.5856	0.5563
	30	0.7673	0.8495	0.9861	0.9669	0.3676	0.5535	0.6059	0.5732
	40	0.7848	0.8611	0.9831	0.9609	0.3708	0.5542	0.5662	0.5362
	50	0.8041	0.8656	0.9832	0.9614	0.3909	0.5737	0.5022	0.5123

2.4.2 Validation based on distance preservation

The comparative distances between samples, as seen in the hyper-space encompassed by the original features, should be preserved by a decent dimension reduction technique. This has been regarded as a crucial characteristic of approaches created for single cells [30]. We calculated the Euclidean distance between every pair of data points, *i* and *j*, in both the original and reduced feature spaces in order to ascertain the same. Two distance matrices, O_{ij} and R_{ij}, were created as a result, and each cell in these matrices stands for the distance between the ith and jth samples. We performed Spearman correlation to determine the correlation between the two matrices, O_{ij} and R_{ij}. A stronger correlation suggests that the distance among samples in the reduced space is either preserved or comparable to the distance among samples in the original space.

For each of the four datasets, SelfE has effectively maintained the sample distance in the condensed feature space, as shown Figures 2.10 - 2.13. However, Kernel SelfE does not show any significant improvement in the obtained correlation when compared to SelfE (Table 2.2).

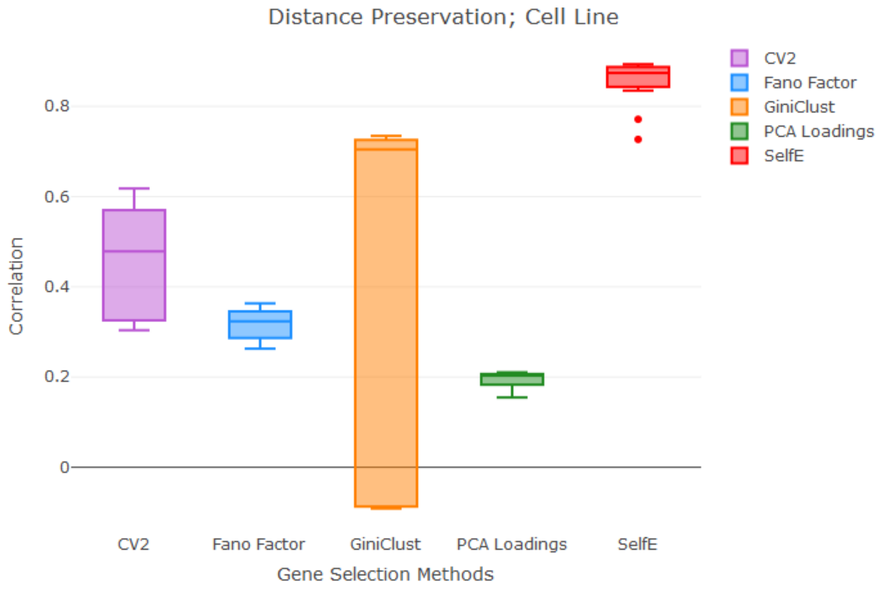

Figure 2.10: Each box displays the spectrum of correlation obtained using various techniques utilizing cell line data. To determine if the distance between points in the original and compressed space is preserved or not, correlation is used.

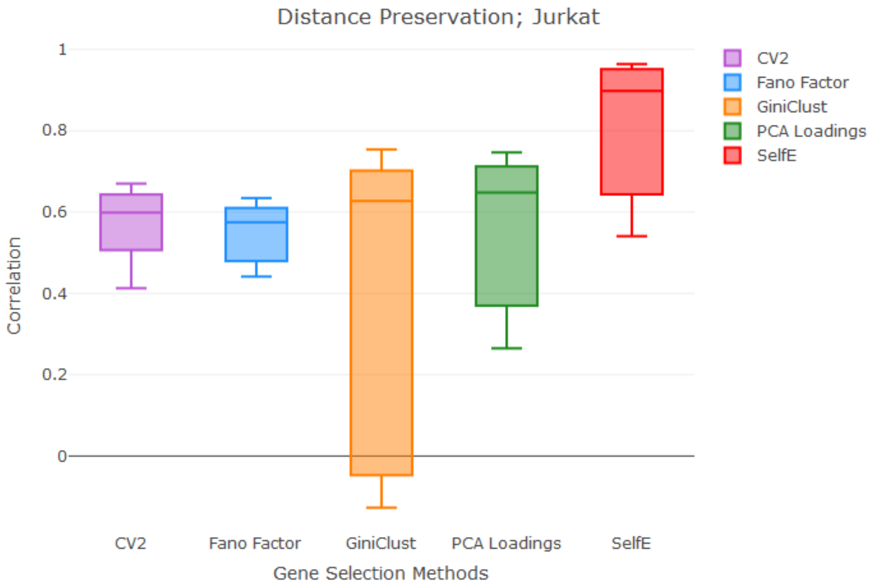

Figure 2.11: Each box displays the spectrum of correlation obtained using various techniques utilizing jurkat data. To determine if the distance between points in the original and compressed space is preserved or not, correlation is used.

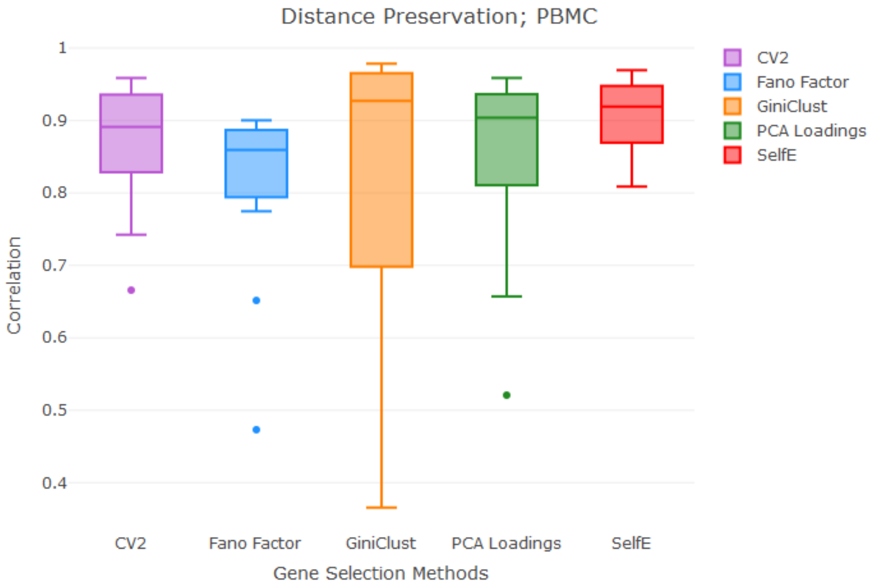

Figure 2.12: Each box displays the spectrum of correlation obtained using various techniques utilizing PBMC data. To determine if the distance between points in the original and compressed space is preserved or not, correlation is used.

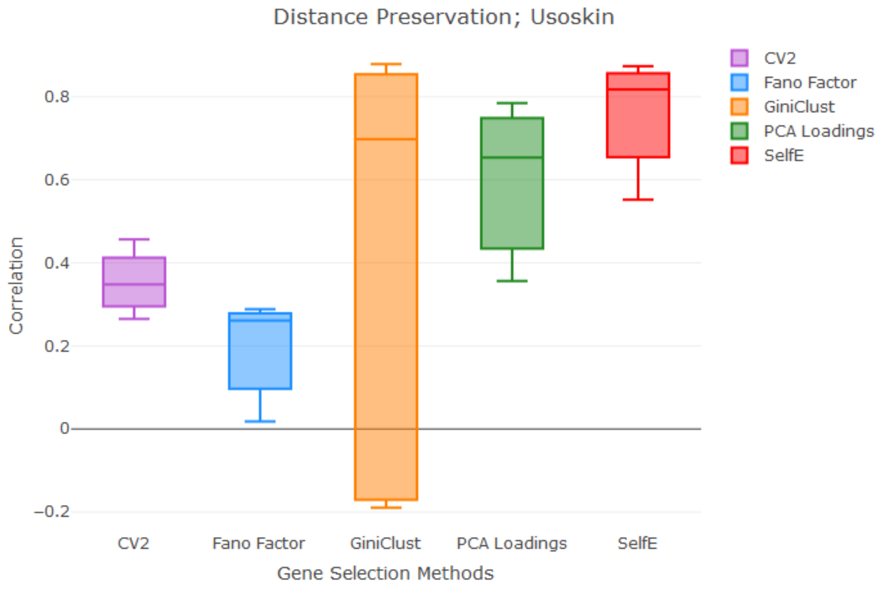

Figure 2.13: Each box displays the spectrum of correlation obtained using various techniques utilizing usoskin data. To determine if the distance between points in the original and compressed space is preserved or not, correlation is used.

Table 2.2: Correlation of the distance between the data points in the original, and the reduced feature space obtained using SelfE and Kernel SelfE

Feature Selection Technique		Datasets			
		Cell Line	Jurkat	PBMC	Usoskin
		Spearman Correlation	Spearman Correlation	Spearman Correlation	Spearman Correlation
SelfE	20	**0.7268**	**0.5406**	**0.8088**	**0.5521**
	30	**0.7714**	**0.5766**	**0.8356**	**0.6055**
	40	**0.8350**	**0.6121**	**0.8517**	**0.6260**
	50	**0.8460**	**0.6538**	**0.8753**	**0.6641**
Kernel SelfE	20	0.6498	0.3843	0.8373	0.2488
	30	0.6142	0.3887	0.8370	0.2903
	40	0.5949	0.3881	0.8359	0.3174
	50	0.5978	0.3968	0.8432	0.3373

2.4.3 Validation based on cell visualization

We must transform the *n* dimensional feature space (*n* is the number of features obtained after applying the feature selection algorithm) to two/three dimensional feature space to visualize if the selected features can distinguish between one or more subpopulations of cells or not.

The two most prominent methods for performing dimension reduction are principal component analysis (PCA) [30] and t-distributed stochastic neighbour embedding (t-SNE) [81]. Using t-SNE, the samples are displayed in two dimensions, and similar cell types are depicted using similar colors. As can be seen from Figures 2.14 - 2.17, the 40 genes that were identified using SelfE for the Usoskin dataset clearly distinguish the subpopulations.

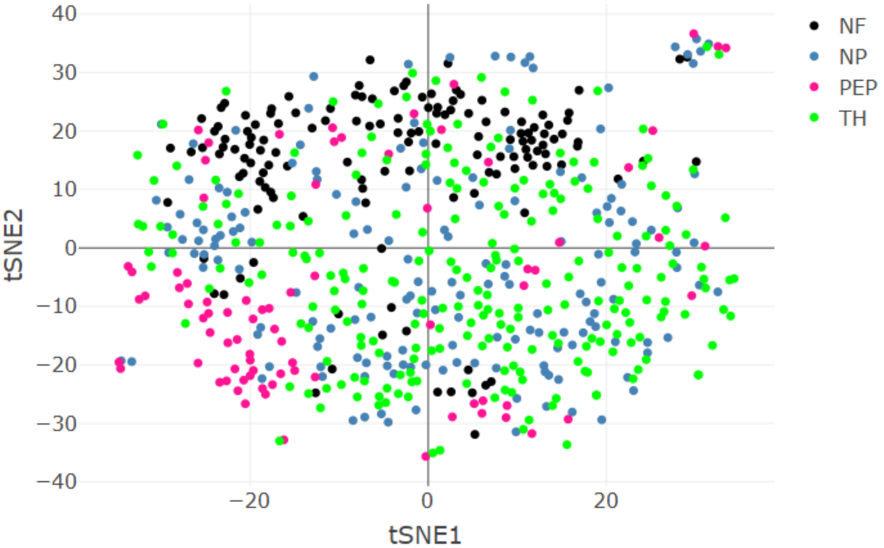

Figure 2.14: The metric coefficient of variance was used to identify the top 40 features from the feature space of the usoskin data. tSNE was then employed on the identified features to check if the different cell types can be distinguished using the selected features or not. The color coding of cells is based on their original annotation.

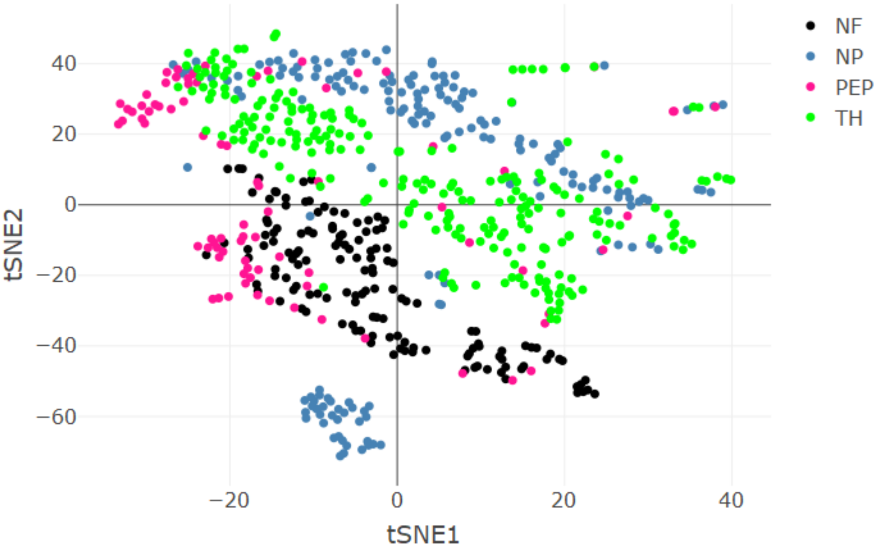

Figure 2.15: The metric Gini coefficient was used to identify the top 40 features from the feature space of the usoskin data. tSNE was then employed on the identified features to check if the different cell types can be distinguished using the selected features or not. The color coding of cells is based on their original annotation.

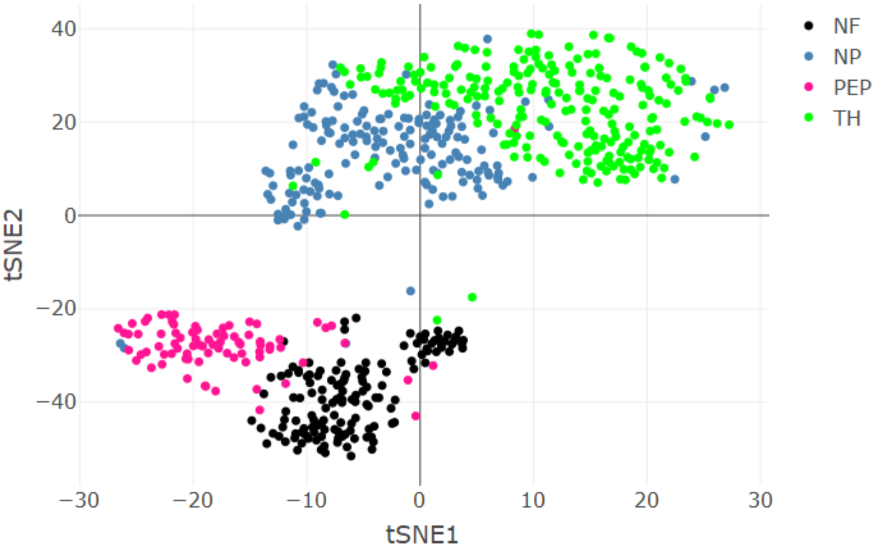

Figure 2.16: The method PCA loadings was used to identify the top 40 features from the feature space of the usoskin data. tSNE was then employed on the identified features to check if the different cell types can be distinguished using the selected features or not. The color coding of cells is based on their original annotation.

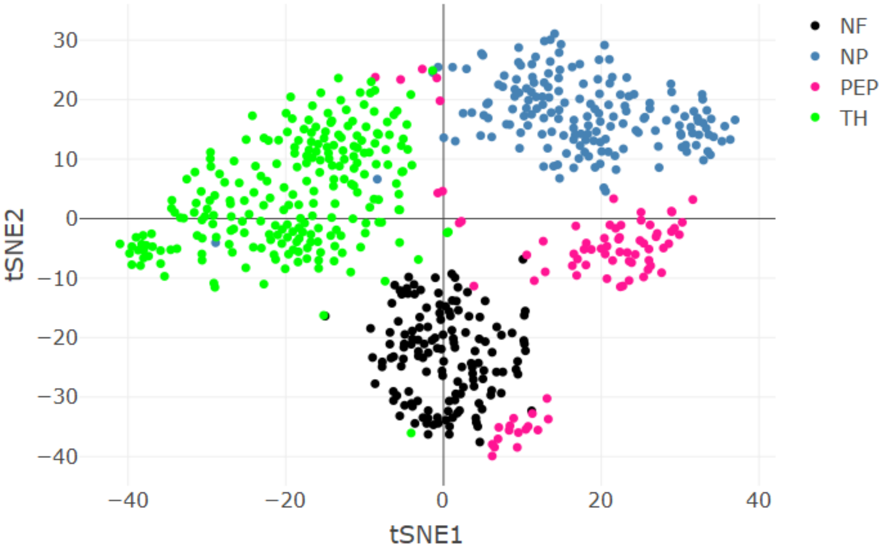

Figure 2.17: The proposed method SelfE was used to identify the top 40 features from the feature space of the usoskin data. tSNE was then employed on the identified features to check if the different cell types can be distinguished using the selected features or not. The color coding of cells is based on their original annotation. As is seen, SelfE has efficiently grouped similar cell types.

2.4.4 Validation based on rare cell identification

The coefficient of variation [59] and Fano factor [26] are two of the current best practice techniques that are used to identify genes that exhibit greater dispersion than other genes. This method raises the possibility of missing useful genes representing uncommon cell groups. On the other hand, self-projection prioritizes the features that can describe the remaining genes while making no distinction across genes based on expression variability. On the other hand, self-projection prioritizes the features that can describe the remaining genes while making no distinction across genes based on expression variability.

We ran a simulation experiment to evaluate SelfE's capacity to detect genes unique to uncommon cell types. We reduced sampled PEP cell line in the Usoskin [40] data such that they made up 5%

of the entire dataset in a controlled study to imitate a small cell population [92]. It should be noted that the data initially had four cell types: NF, NP, PEP, and TH (Dataset Description).

SelfE identified genes *X6330403K07Rik* (*P* = 2.556844e-10, Wilcoxon rank-sum test) and *Tac1* (*P* = 1.802207e-05, Wilcoxon rank-sum test) as biomarkers from the set of 20 features obtained using SelfE (Figures 2.18 - 2.21). The identified genes represent the artificially introduced uncommon cell type PEP, as shown in Figure 2.18. Similarly, Figure 2.19 depicts that artificially planted rare cell type, PEP, can be characterized using genes *X6330403K07Rik* and *Tac1* from the set of 30 features obtained after feature selection. The feature set containing 40 and 50 features after feature selection introduces one more gene, *Calca*, that can be used to characterize rare cell type PEP (Figure 2.20 and Figure 2.21).

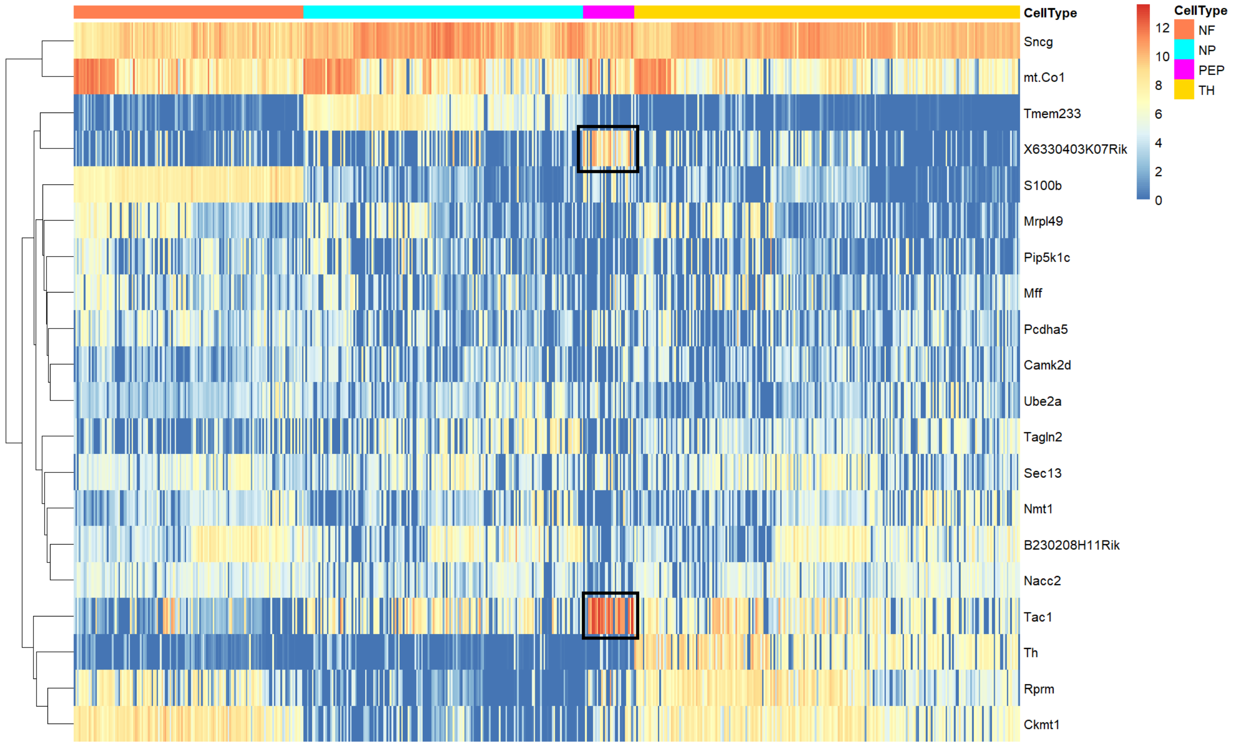

Figure 2.18: For simulated Usoskin data, a heatmap shows the gene expression of 20 genes that were derived utilizing SelfE. The region inside the black box shows how the uncommon cell type PEP expresses marker genes. *Tac1* and *X6330403K07Rik* genes have high expression, making it simple to identify uncommon cells.

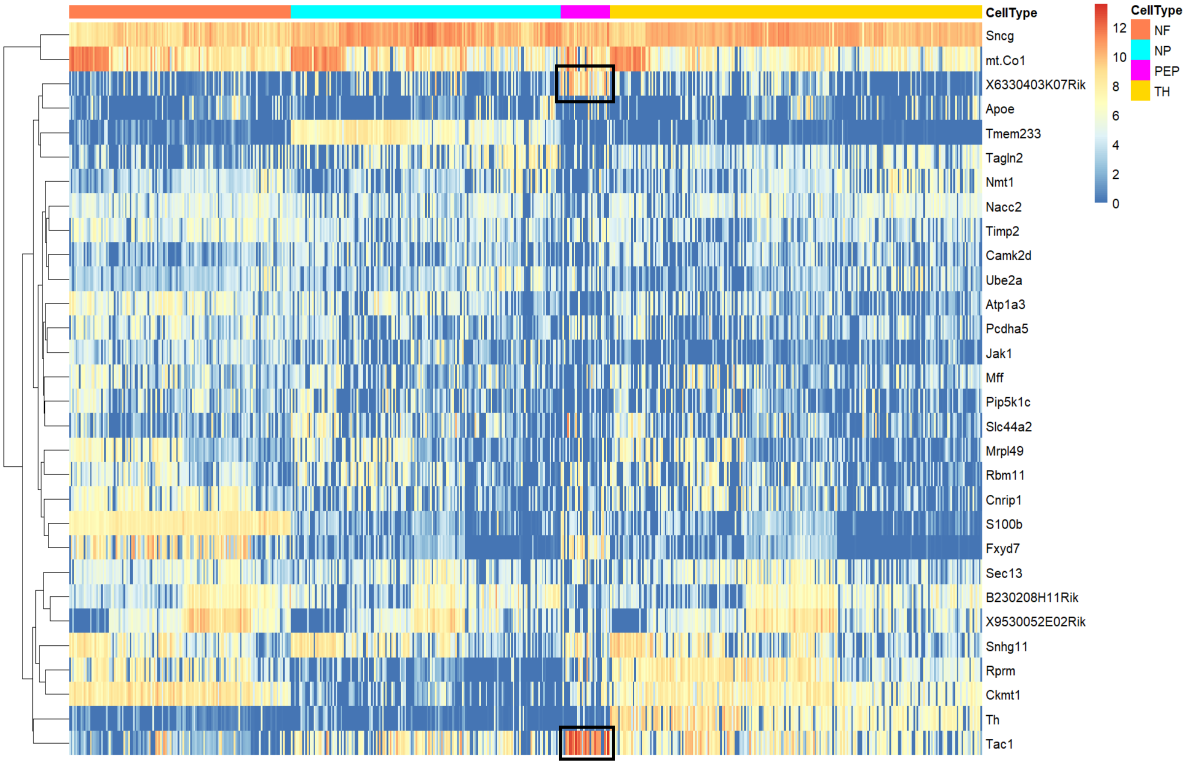

Figure 2.19: For simulated Usoskin data, a heatmap shows the gene expression of 30 genes that were derived utilizing SelfE. The region inside the black box shows how the uncommon cell type PEP expresses marker genes. *Tac1* and *X6330403K07Rik* genes have high expression, making it simple to identify uncommon cells.

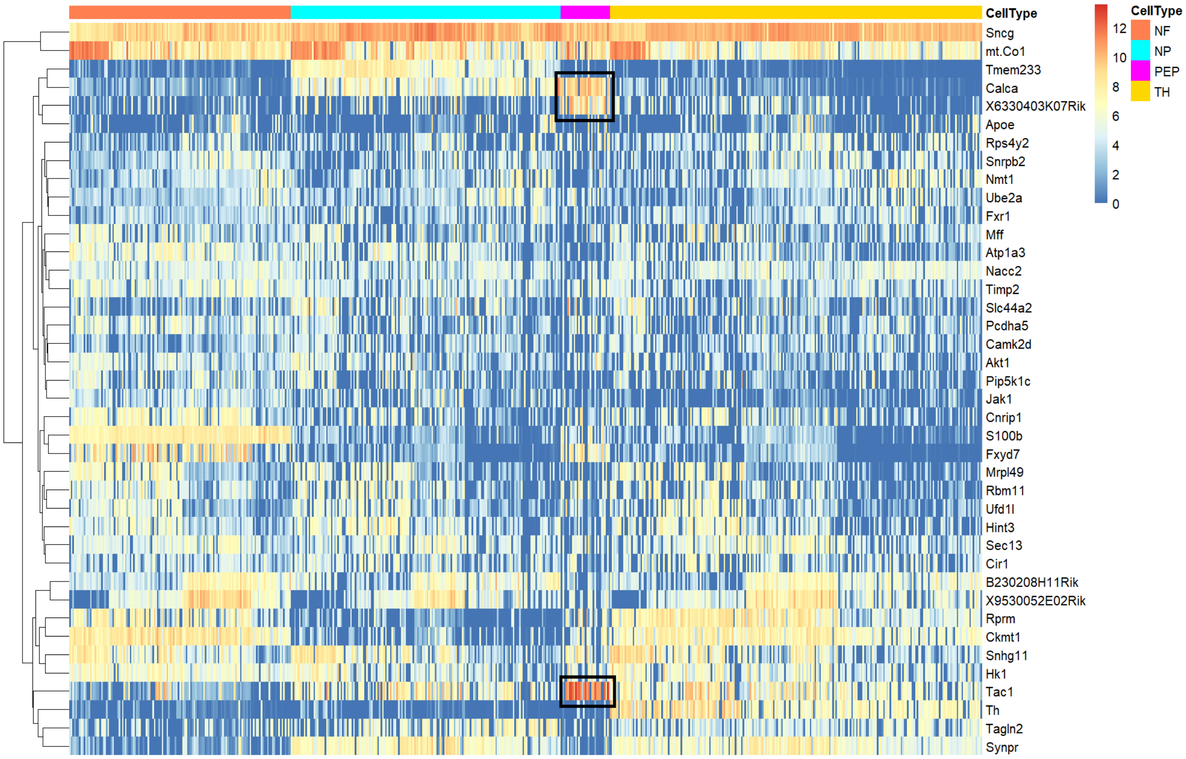

Figure 2.20: For simulated Usoskin data, a heatmap shows the gene expression of 40 genes that were derived utilizing SelfE. The region inside the black box shows how the uncommon cell type PEP expresses marker genes. *Tac1*, *X6330403K07Rik*, and *Calca* genes have high expression, making it simple to identify uncommon cells.

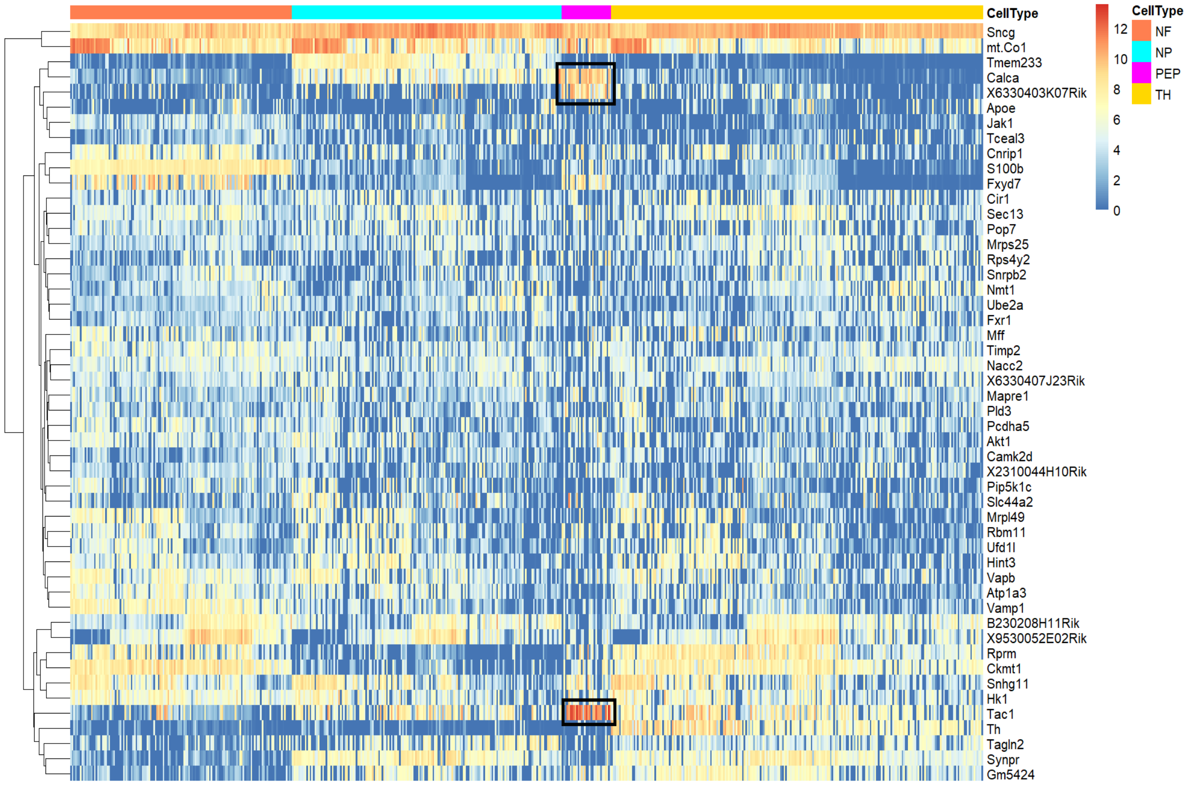

Figure 2.21: For simulated Usoskin data, a heatmap shows the gene expression of 50 genes that were derived utilizing SelfE. The region inside the black box shows how the uncommon cell type PEP expresses marker genes. *Tac1*, *X6330403K07Rik*, and *Calca* genes have high expression, making it simple to identify uncommon cells.

2.5 Conclusion

There is no denying that single-cell RNA sequencing (scRNA-seq) data has a number of advantages, but it also has a number of severe drawbacks, specifically with dimensionality and sparsity. The data obtained through scRNA-seq have a high dimension not only in terms of the number of genes or characteristics but also in terms of the number of cell lines that are sequenced. This not only adds to the amount of time and complexity required to process information, but it also has the potential to cause data over-fitting when training a classifier. The biological noise as well as the intrinsic technical noise that is present during the sequencing process are among the factors that contribute to the sparse character of single cell data [93]. The quantity of RNA that is present in a single cell is extremely low; as a result, it is possible that the expression of a gene will not be recorded, even though that gene is actively expressed in the cell. This results in dropout events, which in turn contribute to the data being sparse. It is crucial to eliminate the unnecessary and redundant features that could conceal a dominant group of features in irrelevant or noisy data [94].

In the literature, various cutting-edge methods for dimension reduction have been proposed. However, clinicians find it challenging to evaluate the condensed feature space that a feature extraction technique provides. This paper made a contribution to the field of feature selection by developing a technique for choosing a minimal gene subset that contains data from the whole feature set. The general principle is to choose genes that are variable in nature. However, the problem with this method is that it is vulnerable to data noise, which has the potential to lead to erroneous alarms for expression variability. In many situations, controlling for average expression might not be enough because there may be many alternative confounders. It should be emphasized that depending on how much weight we assign to the sparsity term, genes that are persistently expressed, such as housekeeping genes, can be represented by just one or a few genes in the reduced feature set generated by SelfE. Therefore, it would be rare for SelfE to invest a sizable number of genes in elucidating genes expressed stably.

This study made a method contribution for feature selection that outperforms full gene sets and standard approaches on the validation measures of clustering accuracy, distance preservation, and cell visualization. SelfE is the most reliable of the evaluated strategies since it consistently retrieves the original annotations with a variety of genes selected. Furthermore, in the reduced feature space, it retains the original distance between cells. In a two-dimensional visualization, the smaller feature set successfully differentiates the cells that belong to different cell types. The rare cell type is successfully distinguished from the population as a whole by the selected attributes. It should be noted that most techniques perform equally well if the number of chosen genes exceeds 200 (Figure 2.2). The proposed feature selection strategy is based on the concept of the union of subspaces. It presumes the samples are grouped on various subspaces and determines the most efficient (in the sense of least squares) representational basis for every subspace from the data. Since it is based on linear algebraic approaches, as opposed to earlier statistical models [26], it is ideologically separate from the present frequently used techniques like the coefficient of variation, PCA loadings, Gini coefficient, and Fano factor. To further enhance its performance, we will test several non-linear kernels in the future.

Chapter 3

Cell identification using cluster aware deep dictionary learning framework

3.1 Introduction

There are numerous reviews [95] on the phenomenon of clustering, which is well recognized. Clustering as a whole examines how originally present clusters within the data form. Perhaps K-means is the most basic and often used method to perform clustering [96]. Samples close to one another (as determined using certain distance metric) are regarded as belonging to the same cluster when using K-means to segment the data by comparative distances. Due to the distance's linear character, the K-means was unable to identify groups that occurred non-linearly. The concept of kernel K-means somewhat addressed this problem [97]. A kernel distance (Gaussian, Laplacian, polynomial, etc.) was defined for segmentation rather than the relative distance between the data points. The kernel K-means and spectral clustering are closely linked [97]. The spectral clustering generalizes kernel distances to any affinity metric and employs graph cuts to divide the groups within the data.

There is a relationship between K-means, kernel K-means, and spectral clustering [98]. Subspace clustering is a fundamentally distinct method [99]. The latter makes the assumption that samples from the same group or cluster will be located within a similar subspace. There are various types of subspace clustering, but sparse subspace clustering (SSC) is the most prevalent [100]. The term "sparse" refers to the assumption made in SSC that the groups only fill a small number of subspaces (out of all available scenarios).

We have so far covered general clustering methods. Identification of a particular cell line is crucial for the subsequent analysis of single-cell data. As a result, clustering is an essential step in the study of single-cell RNA expression data. The amount of mRNA present for a gene in a single cell is measured using single-cell RNA sequencing. However, because there is relatively minimal RNA in a single cell, several genes went undetected even if they are present, and this makes the single cell data sparse. The heterogeneity in the genes that are particular to the cell cycle further complicates this data by introducing biological noise. During an experiment, a significant number of genes are evaluated, but only a small fraction of them are employed for cell-type identification. As a result, single-cell data is of high dimension in terms of features and also consists of features that are redundant in nature. When clustering techniques are directly applied to this high-dimensional data, cell partitioning will not be efficient.

This necessitates the development of tailored strategies. The present strategies for aggregating single-cell data employ established algorithms to extracted or reduced feature sets rather than introducing additional techniques for clustering per se. Seurat [101] uses the ratio of the variance of a gene's expression to its mean for determining highly variable genes. The selected genes are then used to compute a distance matrix of cells or samples. This distance matrix is used to construct the shared-nearest neighbor graph (SNN) which is further passed as an input to the smart local moving (SLM) algorithm for clustering. GiniClust [60] employs the Gini coefficient to select highly variable genes. The Gini coefficient is a mathematical representation of the statistical dispersion, which is computed using a Lorenz curve. In the Lorenz curve, the cumulative percentage of cells is represented on the x-axis, and the cumulative percentage of a gene's expression is represented on the y-axis. In the idle case, if each cell has the same amount of mRNA expression corresponding to a gene, then we will get an upward sloping linear line that represents the perfect equality, i.e., 1. To compute the Gini coefficient for each gene, a ratio of the area between the line of equality and Lorenz curve to the total area beneath the line of perfect equality is taken into consideration. Genes with high Gini index values are selected. A density-based clustering algorithm is employed on the selected genes to determine the clusters in a dataset. Single cell consensus clustering (SC3) [54] uses a cluster-based similarity partitioning algorithm (CSPA) to identify the cluster of cells in the dataset. In SC3, principal component analysis is applied to a cell-cell distance matrix to get eigenvectors corresponding to the top 15 eigenvalues. K-means clustering is carried out on the eigenvectors. A similarity matrix is computed based on the output of the K-means algorithm. If two cells belong to the same cluster, then their similarity will be 1, otherwise 0. In the end, an average of the similarity matrices generated from each iteration of K-means is used to form the consensus matrix. Hierarchical clustering is applied to the consensus matrix to determine the similar type of cells in the dataset.

Today, deep learning has proven successful in every area like image coloring, image captioning, and healthcare. There are just a small number of foundational articles on deep dictionary learning-based clustering, which is significant to note because progress has been mostly driven by supervised tasks [102]. A novel paradigm for deep learning is deep dictionary learning. It has already been applied to supervised classification [103], unsupervised feature extraction [104], and even domain adaptation [105]. However, clustering has never been done with it. It would be a pioneering piece on the subject. Deep dictionary learning has the benefit of being easily adaptable to various cost functions and being mathematically versatile. In this study, we suggest adding K-means clustering and sparse subspace clustering as costs to the deep dictionary learning framework that is unsupervised.

3.2 Proposed formulation

Convolutional neural networks (CNN), stacked autoencoders (SAE), and deep belief networks are the three foundational techniques of deep learning (DBN). Since CNN can only address naturally occurring signals with local associations, the debate on CNN is not pertinent to this situation. Additionally, they are unable to work in an unsupervised manner, so they cannot be a candidate for our topic of interest. We have employed stacked autoencoders for our purpose (deep learning-based clustering). SAE is susceptible to overfitting since it requires learning twice as many parameters (encoder and decoder) in comparison to other typical neural networks.

However, it is easy to manage SAEs in terms of operations and has significant numerical adaptability. Whereas DBN does not overfit since it learns the right number of parameters. But, the DBN cost function cannot be numerically altered.

The finest of both worlds are preserved by deep dictionary learning. It learns the ideal amount of parameters, similar to a DBN, and its cost function is mathematically versatile, allowing it to manage many forms of penalties. This is the main justification behind our decision to base our clustering algorithm on the deep dictionary learning (DDL) architecture. We will include clustering penalties to regularise the DDL cost function in our suggested formulation. The given data is represented with X, where samples or single cells are in columns, and features or genes are in rows. The learned dictionary is represented using D, and it is used to synthesize the data from learned coefficients, which are represented using Z.

$$\min_{D_1,\ldots D_N,Z} \left\| X - D_1\varphi\left(D_2\varphi(\ldots\varphi(D_N Z))\right) \right\|_F^2 \tag{3.1}$$

With K-means, the first clustering penalty will be applied.

$$\begin{aligned} &\min_{D_1,D_2,D_3,Z,H} \underbrace{\left\| X - D_1 D_2 D_3 Z \right\|_F^2 \text{ s.t. } D_2 D_3 Z \geq 0, D_3 Z \geq 0, Z \geq 0}_{\text{Dictionary Learning}} \\ &\underbrace{+\left\| Z - ZH^T\left(HH^T\right)^{-1} H \right\|_F^2 \text{ s.t. } h_{ij} \in \{0,1\} \text{ and } \sum_j h_{ij} = 1}_{\text{K-means}} \end{aligned} \tag{3.2}$$

The cost function for deep dictionary learning has changed. By integrating positivity constraints, we are utilizing the ReLU type cost function in place of activation functions like sigmoid or tanh.

ReLU is preferred over alternatives because of its superior function approximation capabilities [106]. Appropriate changes have been made to the notations in the K-means clustering penalty.

In our case, we will solve problems using the greedy strategy (3.2). In the second layer of deep dictionary learning, we will substitute Z_1 with D_2D_3Z. This results in the greedy approach of the initial level of dictionary learning.

$$\min_{D_1,Z_1} \left\| X - D_1 Z_1 \right\|_F^2 \text{ s.t. } Z_1 \geq 0 \tag{3.3}$$

The output of the first layer of dictionary learning serves as the input for the second layer, that is, Z_1. The equivalent is $Z_2 = D_3Z$. This results in the subsequent equation:

$$\min_{D_2,Z_2} \left\| Z_1 - D_2 Z_2 \right\|_F^2 \text{ s.t. } Z_2 \geq 0 \tag{3.4}$$

No replacement is required for the third and final layer; simply, the result from the second layer is supplied to it.

$$\min_{D_3,Z} \left\| Z_2 - D_3 Z \right\|_F^2 \text{ s.t. } Z \geq 0 \tag{3.5}$$

To solve equations (3.3) - (3.5), we have used algorithm multiplicative updates of non-negative matrix factorization [107]. There is no limit to the number of levels that can be added to this structure.

The last layer's coefficients are used as the input for K-means clustering (Z). This is depicted as:

$$\min_{H} \left\| Z - ZH^{T}\left(HH^{T}\right)^{-1} H \right\|_{F}^{2} \text{ s.t. } h_{ij} \in \{0,1\} \text{ and } \sum_{j} h_{ij} = 1 \tag{3.6}$$

It is addressed using the common K-means clustering approach.

The above explained approach is used to solve the algorithm where K-means is applied on the output of deep dictionary learning approach for clustering. We cannot say that this is the ideal solution since it is greedy as the estimation from the output layer is not transferred to the hidden layers. But there are established solutions for equations (3.3) - (3.6) that we need to resolve.

Apart from the K-means algorithm, we have also applied the sparse subspace clustering approach to the final output of the deep dictionary architecture to perform clustering which is explained in the subsequent lines.

$$\begin{aligned} &\min_{D_1, D_2, D_3, Z, C} \underbrace{\left\| X - D_1 D_2 D_3 Z \right\|_F^2 \text{ s.t. } D_2 D_3 Z \geq 0, D_3 Z \geq 0, Z \geq 0}_{\text{Dictionary Learning}} \\ &+ \underbrace{\sum_{i} \left\| z_i - Z_{i^c} c_i \right\|_2^2 + \left\| c_i \right\|_1, \forall i \text{ in } \{1, \ldots, n\}}_{\text{Sparse Subspace Clustering}} \end{aligned} \tag{3.7}$$

The approach to the deep dictionary learning problem is unchanged. We can resolve it greedily using equations (3.3) - (3.5). The coefficients (Z) from the output layer of the deep learning architecture are given as input to the sparse subspace clustering algorithm. This can be represented as:

$$\min_{c_i's} \sum_{i} \left\| z_i - \mathrm{Z}_{i^c} c_i \right\|_2^2 + \left\| c_i \right\|_1 , \forall i \text{ in } \{1,...,n\} \tag{3.8}$$

After (3.8) has been resolved, the affinity matrix is constructed and then utilized to partition the data employing Normalized Cuts.

3.3 Datasets used for validating the proposed clustering approach

To examine the efficacy of the proposed technique, seven single-cell datasets from various studies were utilized.

3.3.1 Blakeley: The dataset includes three human blastocyst cell lines that were collected using single-cell RNA sequencing (scRNA-seq). The human embryonic scRNA-seq data provided insight into early embryogenesis and was substantiated using protein levels. Thirty transcriptomes from three different cell lines - human pluripotent epiblasts (EPI), extraembryonic trophectoderm cells, and primitive endoderm cells - were used in the study [108].

3.3.2 Cell Line: Fluidigm, a procedure based on microfluidic technology, was utilized to perform scRNA-seq on 630 single cells extracted from seven cell lines. As separate sequencing was done for every cell line, therefore original annotations were linked topically. Nine distinct cell lines are produced by the sequencing - A549, GM12878 B1, GM12878 B2, H1 B1, H1 B2, H1437, HCT116, IMR90, and K562. Two distinct batches of the cell lines GM12878 and H1 were generated [78].

3.3.3 Jurkat - 293T: This particular collection includes 3,300 transcriptomes taken from two distinct cell lines, namely, Jurkat and 293T cells. The transcriptomes are mixed in an artificial environment in quantities that are equal to one another (50:50). The variations and levels of the cell-type-specific identifiers CD3D and XIST are used to label all transcriptomes [89].

3.3.4 Kolodziejczyk: This article describes the single cell study of approximately 704 embryonic stem cells from mouse (mESCs). The cells are grown under three distinct conditions: serum, 2i, and a2i, an alternate ground state. The various cell culture conditions result in distinct mRNA expression in the cells [109].

3.3.5 PBMC: This dataset is comprised of 68,000 peripheral blood mononuclear cell (PBMC) cell lines from healthy volunteers. According to correlation with purified bulk RNA-Seq data derived from fluorescence activated cell sorting (FACS) of typical PBMC subtypes, they are categorised into 11 typical PBMC subtypes. For this investigation, we randomly chose 100 cells from every identified subtype and preserved the whole cluster if the cell count in the particular cluster was under 100 [89].

3.3.6 Usokin: 799 cell lines from mouse lumbar dorsal root ganglion (DRG) make up this dataset. The cells were grouped by the authors using an unsupervised method. 622 cells out of 799 were categorised as neurons, 68 cells had an uncertain classification, and 109 cells were annotated as non-neuronal. On the basis of prominent cell biomarkers, the 622 mouse neuron cells were further categorised into four primary groups: neurofilament containing (NF), non-peptidergic nociceptors (NP), peptidergic nociceptors (PEP), and tyrosine hydroxylase containing (TH) [40].

3.3.7 Zygote: The samples used for the RNA sequencing were taken from 265 individual cells obtained from mouse preimplantation embryos. It includes gene expression levels for cells at various developmental stages, including the zygote, early 2-cell stage, middle 2-cell stage, late 2-cell stage, 4-cell stage, 8-cell stage, and 16-cell stage, as well as the early blastocyst, middle blastocyst, and late blastocyst stages [110].

3.4 Experimental evaluation

In the initial round of tests, we contrasted the suggested algorithm with the two cutting-edge deep learning methods (Table 3.1). The first approach is a stacked autoencoder (SAE) with two hidden layers. The number of neurons that are present in the first hidden layer of SAE is equal to 20, and the number of cell types that are present in the single-cell data is equivalent to the number of nodes that are present in the second layer. A deep belief network is the second technique utilized as a benchmark (DBN). Similar to SAE, DBN has two hidden layers, the first of which has 100 nodes and the second of which has the same number of nodes as the number of class labels in the particular dataset. Twenty nodes made up the first layer of our suggested deep dictionary learning (DDL), and the second layer's nodes are equal to the number of clusters (similar to the arrangement of SAE). The aforementioned setups produced the best results for the SAE and DBN. The K-means algorithm is used for the deepest layer of attributes in both state-of-the-art approaches and the suggested method to identify the clusters in the data.

Table 3.1: Clustering accuracy of the proposed method and existing deep learning techniques on single-cell datasets

Method	Metric	Blakeley	Cell line	Jurkat	Kolodziejczyk	PBMC	Usoskin	Zygote
DBN	NMI	0.190	0.567	0.001	0.032	0.273	0.015	0.385
	ARI	0.056	0.430	0.001	0.171	0.103	0.007	0.296
SAE	NMI	0.181	0.099	0.925	0.170	0.573	0.040	0.107
	ARI	0.011	0.007	0.958	0.215	0.377	0.001	0.006
Proposed Method	NMI	**0.933**	**0.873**	**0.974**	**0.694**	**0.546**	**0.647**	**0.639**
	ARI	**0.891**	**0.801**	**0.989**	**0.645**	**0.359**	**0.642**	**0.359**

Since the original label (class) of each sample or cell line is known a priori, we used two clustering metrics: adjusted rand index (ARI) and normalized mutual information (NMI) to compare how well SAE, DBN, and the proposed method can separate different transcriptomes using the corresponding deepest layer of attributes. We can observe that the suggested approach significantly outperforms the state-of-the-art deep learning technologies. The SAE outcomes are only closely followed by those for PBMC (Figure 3.1 and Figure 3.2).

We employed GiniClust [60], Seurat [111], and SC3 [54], three well-known single-cell clustering approaches, as benchmark methods in the following series of tests (Table 3.2). The setting stays the same for both of our suggested approaches; that is, K-means and SSC are applied to features obtained from the deepest layer.

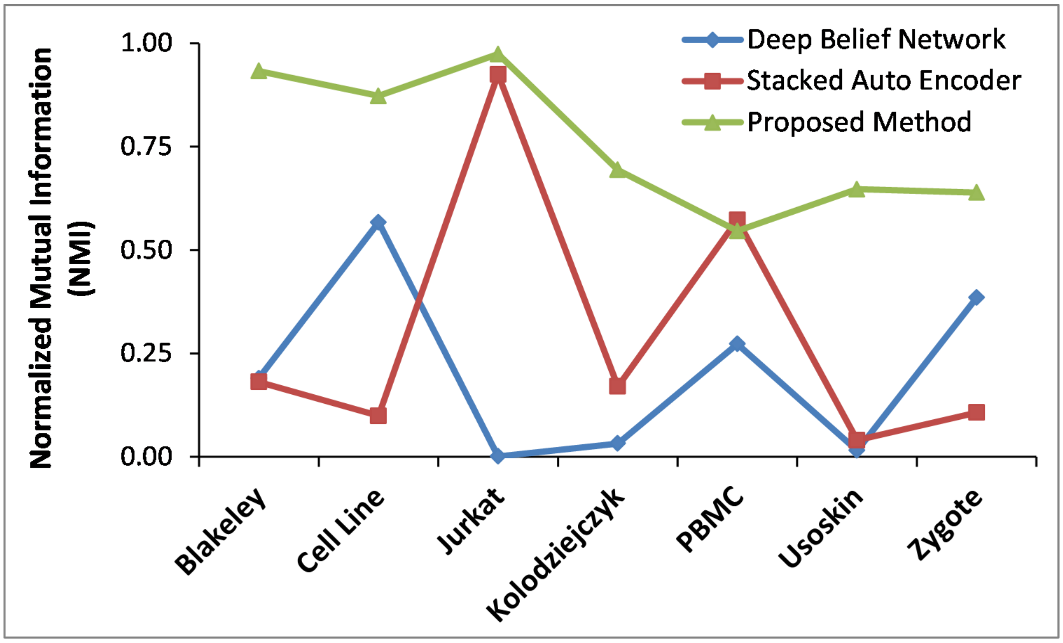

Figure 3.1: The figure depicts clustering accuracy - normalized mutual information (NMI) obtained using state-of-the-art deep learning methods and the proposed clustering algorithm for seven single cell datasets.

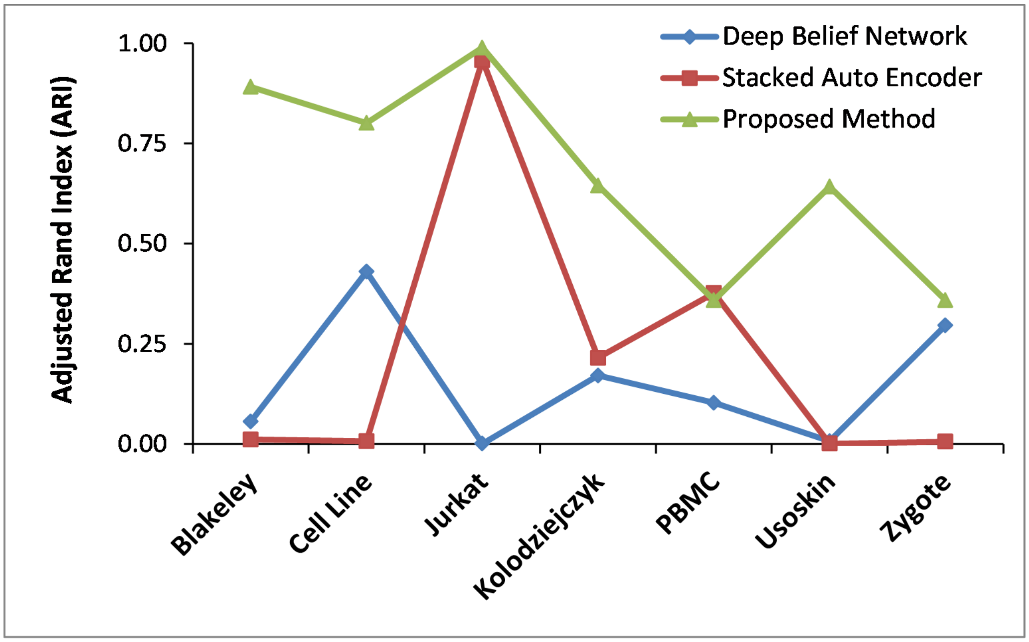

Figure 3.2: The figure depicts clustering accuracy - adjusted rand index (ARI) obtained using state-of-the-art deep learning methods and the proposed clustering algorithm for seven single cell datasets.

Table 3.2: Clustering accuracy of the proposed method and single-cell clustering algorithms on single-cell datasets

Method	Metric	Blakeley	Cell line	Jurkat	Kolodziejczyk	PBMC	Usoskin	Zygote
DBN	NMI	0.277	–	0.007	0.214	0.153	0.061	0.282
	ARI	0.037	–	0	0.055	0.030	0.006	0.025
SAE	NMI	0	0.717	0.946	**0.695**	**0.585**	0.447	0.453
	ARI	0	0.533	0.974	**0.710**	**0.296**	0.382	0.123
SC3	NMI	0.6795	**0.9782**	0.7147	**0.8260**	0.5731	**0.7862**	**0.7115**
	ARI	0.6429	**0.9706**	0.6544	0.6924	0.2834	**0.8677**	**0.4680**
Proposed Method + K-means	NMI	0.933	0.873	**0.974**	**0.694**	**0.545**	**0.647**	**0.639**
	ARI	0.891	0.801	**0.989**	**0.645**	**0.359**	**0.642**	**0.359**
Proposed	NMI	**1**	**0.879**	0.889	0.522	0.481	0.492	0.623

Method + SSC	ARI	**1**	**0.814**	0.821	0.510	0.303	0.453	0.317

For the cell line dataset, GiniClust was unable to produce any clustering outcomes. By using features with a high Gini coefficient value, it achieves clustering. However, the method was unable to locate any highly variable genes and, as a result, was unable to cluster for this specific dataset. In general, GiniClust consistently produces the lowest outcomes. We discover that K-means, out of the two suggested procedures (SSC and K-means), is more reliable and constantly produces positive outcomes. SSC findings are inconsistent, ranging from excellent clustering for Blakely to subpar outcomes for Kolodziejczyk, PBMC, and Usoskin. Just for the Kolodziejczyk and PBMC datasets, Seurat and proposed K-means give comparable results. Seurat is significantly worse than either of our methods for the remaining single-cell datasets (Figure 3.3 and Figure 3.4). However, the performance of SC3 is comparable to the proposed methodologies.

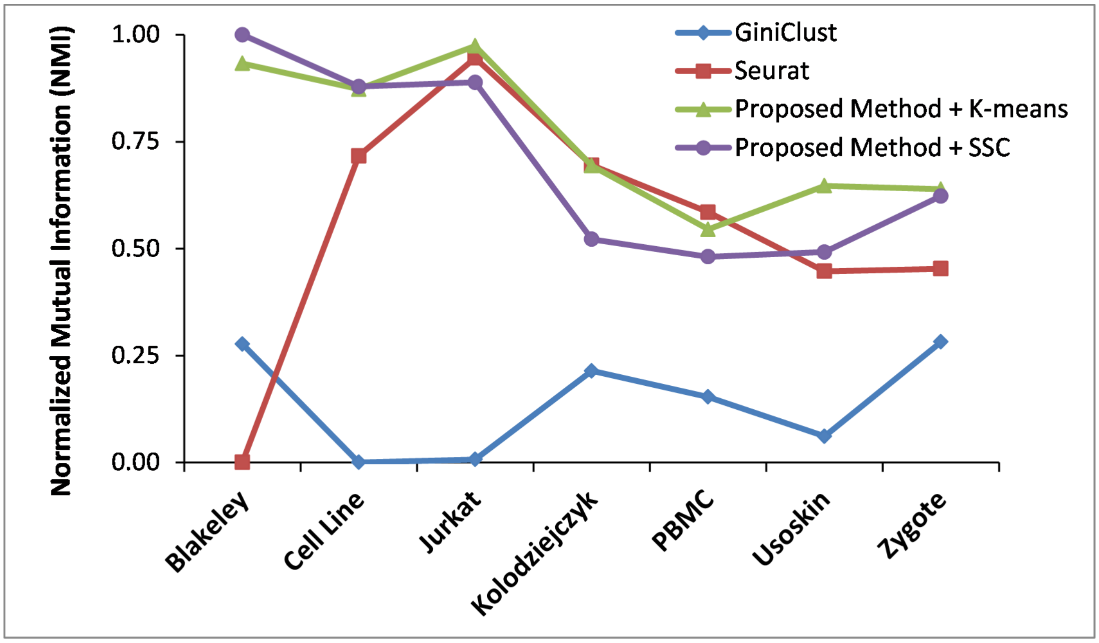

Figure 3.3: The figure depicts clustering accuracy - normalized mutual information (NMI) obtained using state-of-the-art single cell clustering methods and the proposed clustering algorithms for seven single cell datasets.

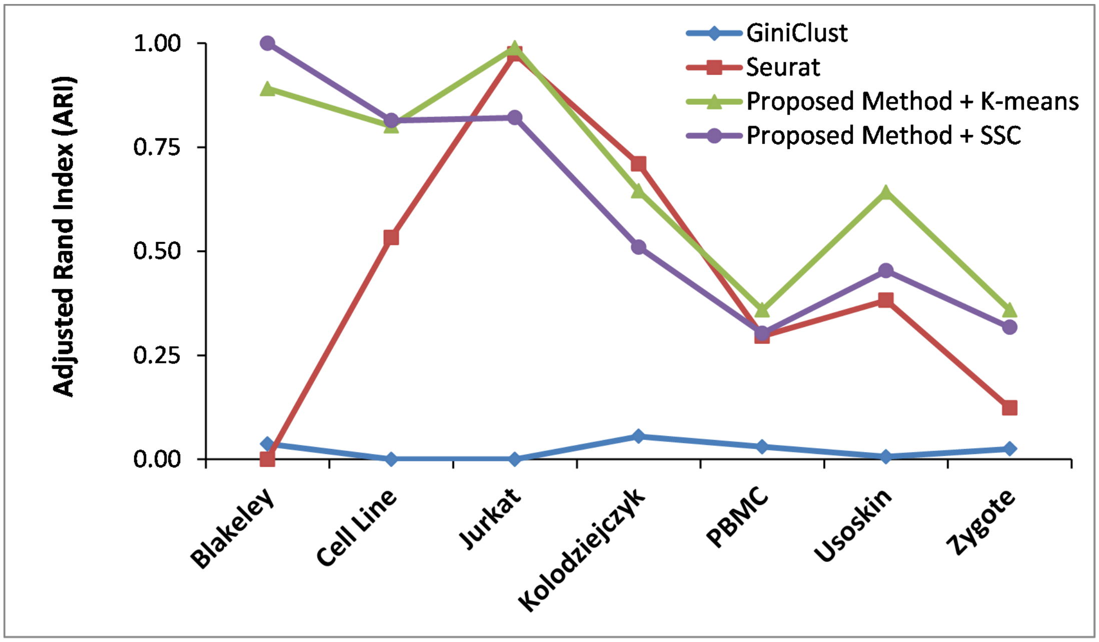

Figure 3.4: The figure depicts clustering accuracy - adjusted rand index (ARI) obtained using state-of-the-art single cell clustering methods and the proposed clustering algorithms for seven single cell datasets.

3.5 Conclusion

A deep dictionary learning-based clustering approach is suggested in this paper. It creates a low-dimensional embedding of the input single cell data (where cells are present in rows and genes are present in columns); this embedding is then given as input to a clustering algorithm. Each sample is represented by a low-dimensional embedding, which is trained in a way that causes the output to be intuitively clustered.

We compared the suggested approach to cutting-edge deep learning approaches (SAE and DBN) and customized single-cell RNA clustering methods in order to assess its performance (GiniClust and Seurat). In terms of clustering accuracy, ARI and NMI, the proposed deep dictionary learning-based clustering methodology performs better than the benchmark methods on seven different single cell datasets.

There is no communication between the deep and shallow layers in the existing method, which makes the current method greedy and, therefore, sub-optimal. Future work will involve employing cutting-edge optimization technologies to jointly solve the entire formulations (3.2) and (3.7).

Chapter 4

Literature mining discerns latent disease-gene relationships

4.1 Introduction

Since the start of the genomic era, there has been a significant increase in the count of released biomedical publications. In peer-reviewed journals, approximately 3,000 submissions are published each day [63]. This has led to a vast body of scientific material that far exceeds the capacity of human comprehension. Considering the multiplexity of human biological systems, it is almost hard to find a qualitative and quantitative overview of the pathogenic roles of genes across multiple organs and tissues by conducting a web search of the available research. A lot of community-level work is being done now to map out the functionally diverse cellular subtypes in tissues of concern in order to build a molecular atlas of normal human organs. Due to the abundance of variables and the heterogeneity of illness states, such an attempt is rare in disease biology. Disease-gene association knowledge-based archiving is carefully and manually curated in the vast majority of cases. In this context, it's important to mention DisGeNET [112] and Online Mendelian Inheritance in Man (OMIM®) [113]. An automated variant of manual data curation, disease-gene relationship retrieval from biomedical corpus has been the main focus of NLP-based efforts in this area [114]. To extract disease-gene connections, Quan and colleagues employed Probabilistic Context-Free Grammars [115], whereas Zhou et al. used a straightforward technique based on term co-occurrence to retrieve the same [116].

The recent development of words with distributed representation has revolutionized text mining. Continuous embedding of words often uses effective neural networks to absorb millions of pages and learn word representations. At Google LLC, Mikolov and coworkers originally presented the concept of learning a continuous bag of words to represent words [64]. Since then, the area has advanced significantly. Recently, models based on transformers, like BERT (bidirectional encoder representations from transformers) [68], have become the cutting edge. Text mining and language modeling have become more popular in recent years in the areas of biological research. Word embedding strategies were used by Gideon et al. to discover correlations between various brain regions [72]. Similar strategies have been applied to both drug discovery [76] and the prediction of compound-protein interactions [117]. BERT utilizes a lot of resources. In the latest project, Lee and colleagues used BERT to analyze massive biomedical literature and provide embeddings for several billions of words. Such generalized word embeddings, according to our theory, might not accurately describe the molecular disease. To get around this, we created a methodical approach to find abstracts from a collection of over ~18 million biological abstracts that indicate associations between the deregulation of specific genes and diseases [118]. The BERT based word embedding model (also known as *PathoBERT*) was then tuned using about 4.5 million carefully chosen abstracts in accordance with NLP guidelines. We only discovered a few research that extracted disease-gene correlations from sentences using word embedding techniques [119]. By design, these techniques are unable to uncover latent knowledge about disease-gene relationships that may not be explicitly mentioned in publications. Interestingly, there is another school of thought in this field that uses node embedding in heterogeneous networks that include genes, diseases, chemicals, etc. The research by Zhou and associates [120] stands out among these since it offers around ~11.2 million possible disease-gene correlations. In a related study, Yang and associates [121] used data on both genes (like gene ontology and protein-protein interactions) and diseases (like symptoms) to enhance the efficacy of disease-gene prediction. A big help in this area is the survey study by Yang and associates on representation learning on heterogeneous networks [122]. However, the adoption of network-based techniques depends on the availability of thorough network topologies. Moreover, it hasn't been shown that these techniques are effective at uncovering new information or representing latent knowledge.

We investigated how well PathoBERT might represent genes, diseases, organs/tissues, cell types, and their interactions in the related embedding space. We compared the best manually curated disease-gene associations with PathoBERT and other embedding approaches. The performance of PathoBERT greatly outperformed that of the available ready-to-implement embeddings. In order to visualize the tissue-specific pathogenic roles of genes as a heatmap on a human body architecture, we have created the R package Pathomap. Furthermore, we were able to extract latent knowledge outside the scope of the training literature corpus owing to these disease-specific word embeddings.

4.2 Methods

4.2.1 Dataset description

Approximately 18 million abstracts from PubMed that had been released up until May 2019 were downloaded [118]. Abstracts, according to our hypothesis, encapsulate the core of the entire manuscript [123]. Additionally, many articles are inaccessible for mining because of paywalls.

4.2.2 Manual annotation of abstracts

We manually selected a significant subset of abstracts that serve as positive and negative examples, respectively, in order to train a classifier that autonomously distinguishes abstracts related to disease-gene (patho-abstracts) from irrelevant ones. For this, we searched for literature in the fields of genetics and molecular biology. The NHGRI-EBI GWAS Catalog [124], COSMIC (Catalog Of Somatic Mutations In Cancer) [125], and OMIM (Online Mendelian Inheritance in Man) databases were our main sources of manuscripts. As a result, two sets of abstracts were produced: one that describes the direct functions of genes in the pathogenesis of diseases and the other that does not reference apparent disease-gene links. The remaining abstracts that did not include disease-gene linkages were categorized as ‘non-relevant’ or ‘negative’ data, while the abstracts that did indicate gene-disease relationships were referred to as ‘relevant’ or ‘positive’ data. The number of abstracts present in the group of pathological abstracts is 1,412, while the group of non-pathological abstracts consists of 687 abstracts. A team of three annotators - co-authors of this paper - annotated each abstract as ‘pathological’ or ‘non-pathological.’ Then, for each pair of annotators, we calculated the alpha-reliability inter-annotator agreement 1u [125,126], $\alpha_{AB} = 0.83$, $\alpha_{AC} = 0.84$, $\alpha_{BC} = 0.99$. The majority vote process was used to determine the final annotations.

4.2.3 Stratifying pathological abstracts from non-pathological abstracts

To narrow down the abstracts that illustrate the pathogenic significance of genes, an abstract stratification approach was created. As seed data, we used pathological and non-pathological sets of abstracts. These abstracts were subsequently separated into 75:25 train and test sets. Each abstract underwent a standard set of preprocessing operations, including lowercasing the entire text and eliminating stop words and punctuation. Additionally, we use three techniques to get a word embedding for each abstract:

- **Skip-gram:** To learn a vector representation of each word in the vocabulary, we trained the model for 1000 epochs and generated a vector of dimension 700 using Word2vec's basic skip-gram architecture. Here, the window size is 5. The embedding for each abstract is obtained by taking the average of all the word vectors found in the document.

- **BioSentVec:** Since the model was already optimized for biological text, we used it here without further tweaking. The model received each abstract we provided in natural language and produced a 700-dimensional vector [127].

- **BioBERT:** It is a BERT-based model that has been improved using biological information. Each subword is represented in 768 dimensions. In order to obtain the approximation for the entire abstract, we finally took into account the average of these embeddings.

After obtaining an abstract representation using all of the methods mentioned above, we trained ML-based classification models to categorize an abstract into pathological and non-pathological subgroups. All of the classifiers were created using the Sklearn library[128]. We used a grid search to find the ideal hyperparameters, and Table 4.1 presents the findings. The approaches' descriptions and cross-comparison results for the abstract categorization tasks are given below.

- **Support vector machine (SVM):** To establish an association between the representation and the discrete labels that define if the abstract is pathological or not, we trained a three-degree polynomial kernel with balanced class weights [129].

- **Extreme gradient boosting (XGBoost):** To learn the necessary mapping, we utilized 100 estimators with 40 jobs, and a learning rate of 3e-1 [130].

- **Logistic regression (LR):** To provide us with the necessary mapping, the default L2 regularisation with an lbfgs solver converged in about 100 epochs [131].

- **Random forest:** The random forest classification model was learned with a depth range of 2 [132].

- **Multi-layer perceptron:** Within 300 epochs, a straightforward multi-layer perceptron was learned [133].

Table 4.1: The table below illustrates the performance of several classification methods while differentiating abstracts with disease-gene relationship from those of generic type. Finally, the extreme gradient boosting (XGBoost) algorithm was applied as a classifier.

Embedding Type	Model	Accuracy	Precision	Recall	F1	Kappa Score
BioBERT	SVM	92.19	94.39	94.13	94.26	82.02
	XGBoost	92.57	93.93	95.25	94.59	82.73
	LR	90.85	93.53	93.01	93.27	78.99
	MLP	92.57	96.5	92.45	94.43	83.27
	Random Forest	88.95	94.91	88.54	91.61	75.47
Word2vec	SVM	88.95	93.71	89.42	91.52	75.69
	XGBoost	85.33	85.63	93.71	89.49	65.36
	LR	86.66	91.17	88.57	89.85	70.42
	MLP	87.23	91.98	88.57	90.24	71.8
	Random Forest	69.52	98.46	55.14	70.69	43.79
BioSentVec	SVM	92.95	95.36	94	94.67	84.25
	XGBoost	**93.14**	94.6	95.14	**94.87**	**84.52**
	LR	90.09	93.31	91.71	92.5	77.9
	MLP	92.57	94.55	94.28	94.44	83.3
	Random Forest	88.38	98.65	83.71	90.57	75.69

Comparing various classifiers, the XGBoost model delivered the overall best result on the majority of measures (accuracy, recall, and kappa score). Because of a disparity in the data, we used kappa as a leading metric. Therefore, for the classification of ~18 million abstracts, we ultimately employed the XGBoost model learned using the embeddings generated from the BioSentVec model. We used a threshold of 0.8 as a rigorous selection criterion, and as a result, 45,76,952 pathological abstracts were obtained.

4.2.4 Fine-tuned PathoBERT

The 4.5 million pathological abstracts that were obtained following categorization were to be represented by embeddings, and that was our goal. To do this, we first tokenized the data using the NLTK word tokenizer in Python [134]. After that, we eliminated stop words and punctuation from the set of tokens. Lower-casing the articles, as we discovered, reduced the computational burden and the vocabulary size. For all 18 million abstracts, we employed named entities (NEs) from Kim et al. [118] to guarantee meaningful embeddings for biological words. Pubtator [135] served as the foundation for NER for Kim et al. Additionally, we included the names of diseases as NEs from the DisGeNET website [112]. In n-grams connected to the NEs, spaces were swapped out for underscores ("_"). The bigram "adipose tissue," for instance, was changed to "adipose tissue."

We investigated four distinct techniques to obtain the embeddings for these pre-processed abstracts: BioBERT [63], BioSentVec [127], BERT [68], and the skip-gram Word2vec variation [64]. We used publicly available BioBERT and BioSentVec. We used our pre-processed pathological abstracts to fine-tune the BERT-base for the masked language modeling job (source: huggingface repository) for two epochs. We used the average representation from the final four layers of the tuned BERT model (called *PathoBERT*) to obtain the representations connected to a query (word, phrase, sentence, and paragraph). For this, a 32GB Tesla V100 GPU was utilized. To create an embedding for a query term, a Word2vec skip-gram model was trained from the beginning by utilizing the pathological abstracts. Out of the 4.5 million abstracts, only ~1.2 million had species tags that may be related to humans (human, boy, girl, children, man, woman, men, women, patients, and patient). On the basis of these additional filtered abstracts, we fine-tuned *PathoBERT* and Word2vec skip-gram.

4.2.5 Word similarity measurement

The cosine distance between the corresponding word vectors was utilized to gauge the semantically similarity between the two words. Since it assesses how frequently a pair of terms appear together in publications pertaining to pathology, cosine distance in this instance is referred to as a “*Patho-score*.”

4.2.6 Pathomap for visualizing the impact of a gene on different tissues

We developed an R tool that enables users to contrast tissue-wide expression levels from the GTEx gene expression archives with the visualization of a gene's tissue-specific pathogenic involvement [136]. On a human body model, we visualized Patho-scores using the R tool gganatogram [137]. We employed median normalization and log transformation to the GTEx expression matrix in order to visualize gene expression levels.

4.2.7 Monte-Carlo estimation of *P*-value associated with Patho-scores

As word-embedding tools are unsupervised, they are not noise- and spurious-free. Therefore, assessment of statistical significance is essential. We determined P-value

$$P = \frac{\#(PS > r_i)\ \forall i + 1}{|r| + 1} \tag{4.1}$$

where *PS* is the observed *Patho-score* and *r* is a collection of *Patho-scores* calculated between arbitrary word pairs (Tissue-Gene pairs or Disease-Gene pairs depending on the context).

4.2.8 Network diffusion based novel marker discovery in colorectal cancer

Our method involved a genome-wide multiplex network with three different types of connections: disease-disease, disease-gene, and protein-protein [https://academic.oup.com/bioinformatics/article/35/3/497/5055408]. In this type of multiplex network, Random Walk with Restart (RWR) can be described as follows.

$$p_{t+1} = (1 - r)\, M\, p_t + r\, p_0 \tag{4.2}$$

where M is a degree-normalized adjacency matrix, and p_0 is a vector representing the initial probability distribution with non-zero values at the seed nodes. The stationary probability distribution is attained when the difference between the vectors p_{t+1} and p_t is small after a number of iterations. These vectors' elements show the proximity between each graph node and the seed (s). 0.7 is set as the value of restart probability r. Either eigen-decomposition or numerical techniques like the power iteration algorithm can be used to solve an RWR. We conducted RWR using the R package *RandomWalkRestartMH* with seeds “colorectal cancer” and “PLA2G2A” (a gene believed to be related to the disease). With the use of *PathoBERT*, we evaluated a few relatively unstudied proteins that were close to the seed nodes.

4.3 Results

4.3.1 Training a disease-focused language model

There are several ways to derive disease-gene connections using literature mining, as was mentioned in the introduction section. The models are not trained to execute the masked-language-modeling task in the context of diseases, despite the fact that they do a respectable job of automatically finding literature evidence for a disease-gene link. Because of this, they are ineffective in capturing the latent knowledge space encompassing various biological elements in related to diseases. To get around this, we divided the ~18 million abstracts between those that mention disease-gene connections and those that do not. Thus, ~4.5 million *patho-abstracts* were produced. With these filtered abstracts, we refined the underlying BERT and trained skip-gram afresh (including versions completely focused on literature labeled with "human" as the concerned species). Figure 4.1A shows the process flow. The most accurate method for separating patho-abstracts from the rest of the biomedical literature was BioSentVec based embeddings (Figure 4.1B). Benchmarking *PathoBERT*, a fine-tuned version of BERT, against a number of ground-truth notions, produced the best results (described in upcoming sections). We developed a case study to illustrate the semantic similarities between disease-symptom-gene associations in order to assess the meaning of the word vectors. We examined three such triplets: migraine - headache - *ATP1A2* [138], asthma - cough - *CCR2* [139], and neuroblastoma - chest pain - *PHOX2B* [140]. Using Principal Component Analysis (PCA), we plotted the vectors connected to these triplets (Figure 4.1C). Because of the consistent directionalities in the relative arrangement of the words across the triplets, there are constant vector activities among words that reflect ideas like "symptom of" and "causative gene for."

4.3.2 Disease linked genes are uniformly distributed across the human genome

We hypothesized that disease-related genetic hotspots might exist. In order to verify this, we calculated the total number of genes present in each of the continuous 1 million base pair sections of DNA that made up each human chromosome. Additionally, we counted the instances in which each of these stretches' associated gene(s) occurred in the *patho-abstracts*. With Spearman's correlation coefficient of 0.7, we discovered that these values were strongly associated throughout the 1 million base-pair long DNA segments (Figure 4.2).

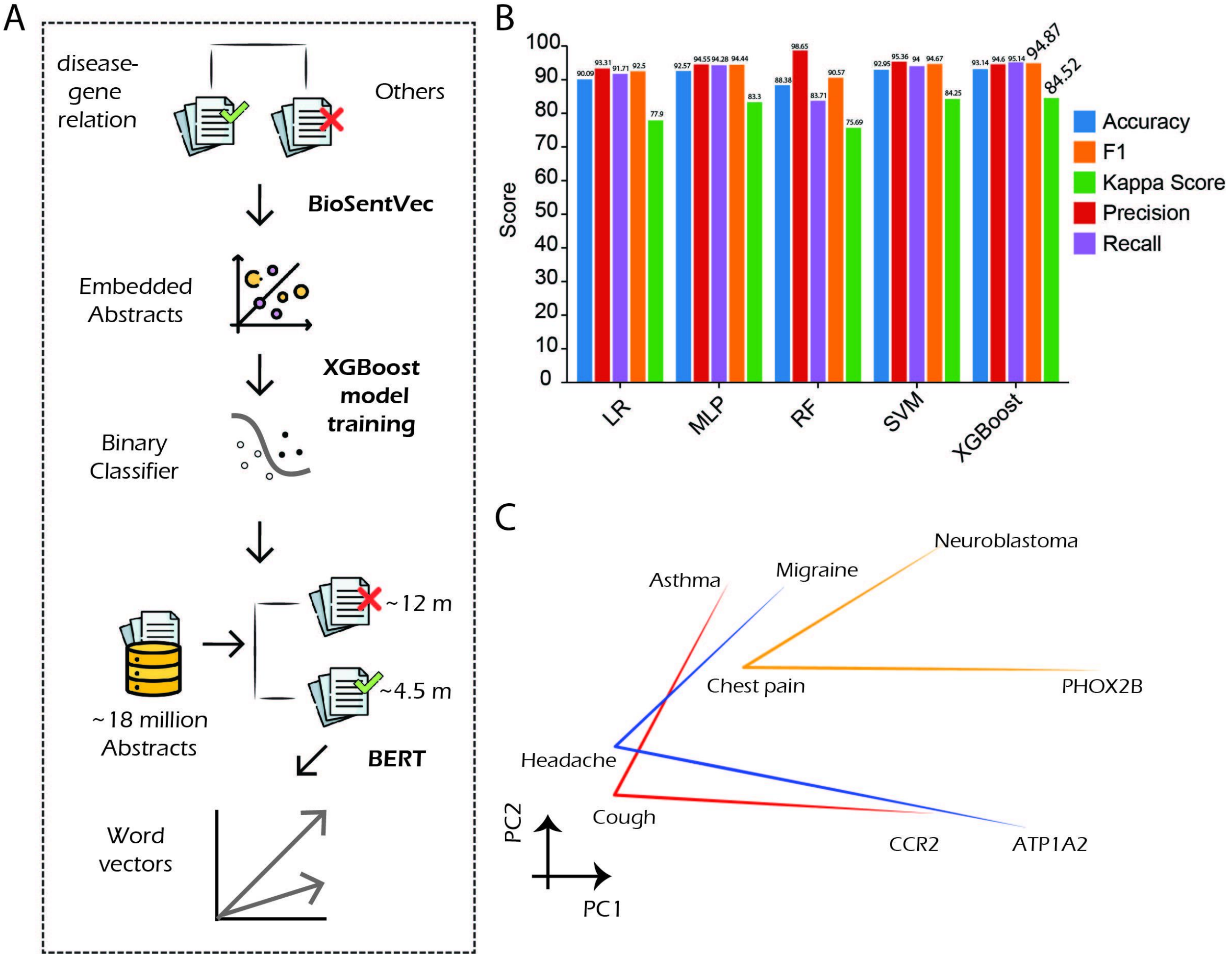

Figure 4.1: (A) The procedure for learning representations from abstracts indicating disease-gene relationships. (B) In recognizing papers describing gene-disease links, BioSentVec offered the best accuracy. (C) The triplets correspond to the disease, symptom, and related gene in the word-vector domain.

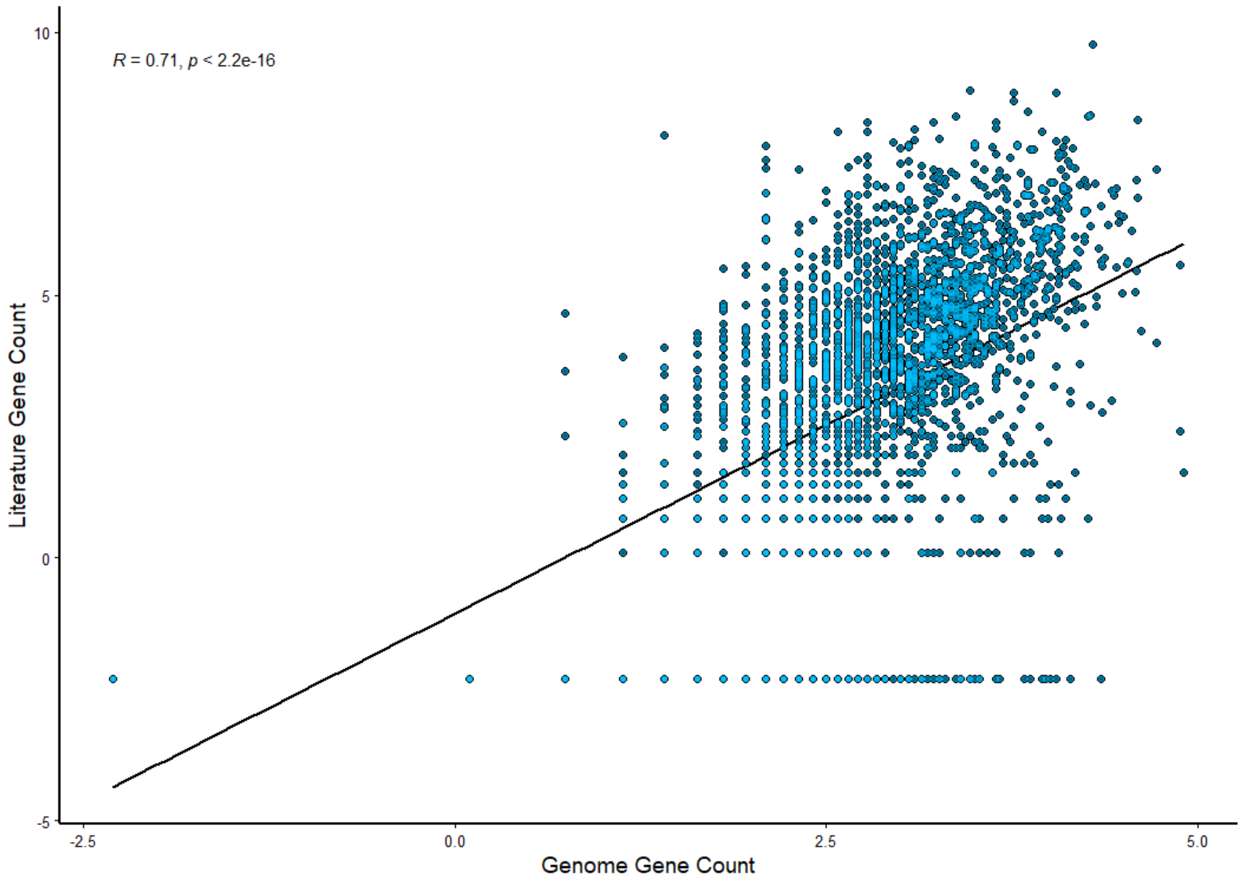

Figure 4.2: Correlation between the number of genes in each of the human DNA's 1 million bp windows and the frequency of those genes in *patho-abstracts*.

4.3.3 PathoBERT maximizes predictability of disease-gene associations

We tested six separate models, as described in the Methods section, to assess how well they could find manually vetted ground-truth disease-gene associations found in DisGeNET [112]. To do this, we created two sets: a positive set (related gene-disease pairs) made up of about 36,000 disease-gene associations (tokens for certain pairs were absent from Word2vec-based embeddings), and a negative set (random disease-gene pairs) that was comparable in size. Using embeddings from several techniques, we calculated the cosine distance between disease-gene combinations. By formulating the identification of the disease-gene relationship as a binary classification problem with cosine distance as the sole explanatory variable, ROC-AUC analysis was carried out. Unexpectedly, only PathoBERT and BioSentVec demonstrated non-random prediction, with PathoBERT taking the lead with an AUC of 0.8 (Figure 4.3). Non-random accuracy was produced using BioBERT and tweaked Word2vec. The performance of BERT/Word2vec trained on human-related literature was equally poor. In addition to providing higher performance, PathoBERT generates word vectors for any word or phrase, even though they are not in the training corpus, thanks to its underlying design. We also contrasted our approach with the DISEASES database, which uses text mining to infer connections between genes and diseases [141]. DisGeNET and the DISEASES database shared 13,740 gene-disease pairs, and *PathoBERT* consists of 35,467 DisGeNET gene-disease pairs with significant *Patho-scores* (P-value < 0.05) (Figure 4.4).

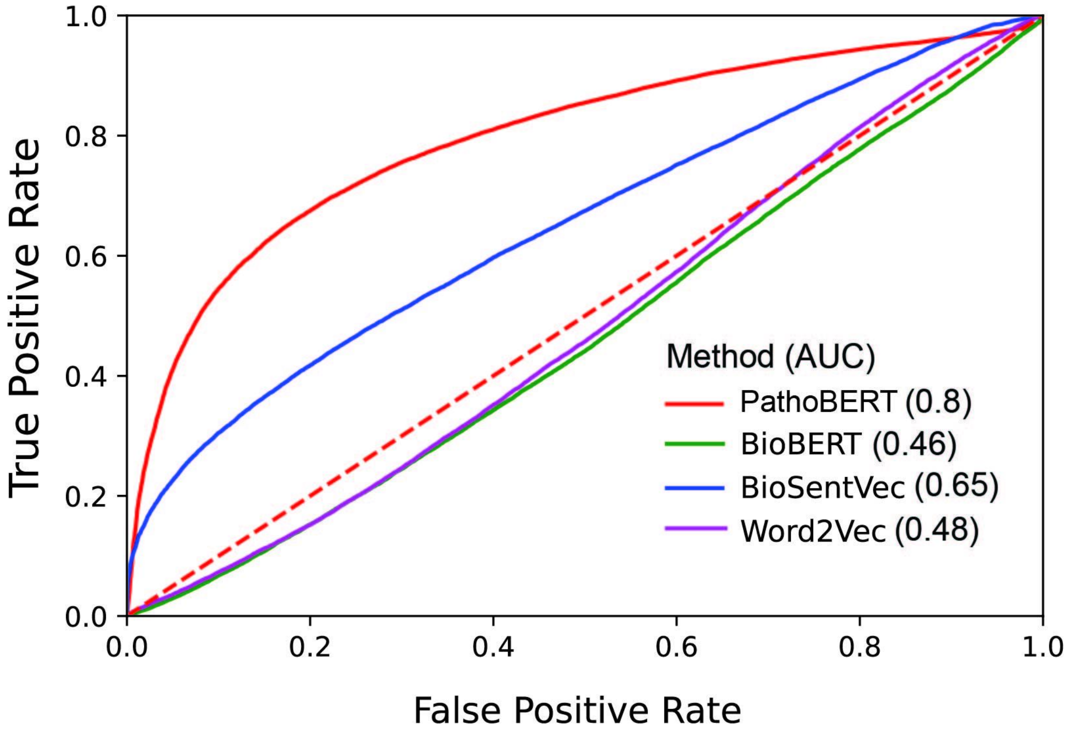

Figure 4.3: ROC plots and related AUC values show how well different embeddings perform in identifying real gene-disease relationships as opposed to spurious ones. For the purpose of this evaluation, the task of finding disease-gene associations is framed as a bi-class classification problem, in which positive associations are those that have been manually archived in the database DisGeNET, and negative associations are produced by pairing genes at random with diseases. Take into account that for the random examples, it is difficult to guarantee the lack of a biologically relevant link.

4.3.4 Gene's pathological role at the resolution of organs and cell-types

Transcriptomic, proteomic, and epigenomic assays have been used in numerous studies since the dawn of the genomic era to build molecular profiles of healthy tissues. These initiatives have been made possible by single cell transcriptomics, which offers phenotypic heterogeneity in cells with undreamt resolution. Recent years have seen a rapid increase in single cell atlasing efforts due to the development of a high throughput single-cell transcriptomics platform, which has helped to untangle the diverse molecular layout of normal tissues. Since the genetic abnormalities that underlie the related pathophysiology of afflicted tissues are non-deterministic, it is challenging to conduct such research on the same. We created the R package *Pathomap*, which displays a heatmap of the pathogenic involvement of genes across organs on a human body template.

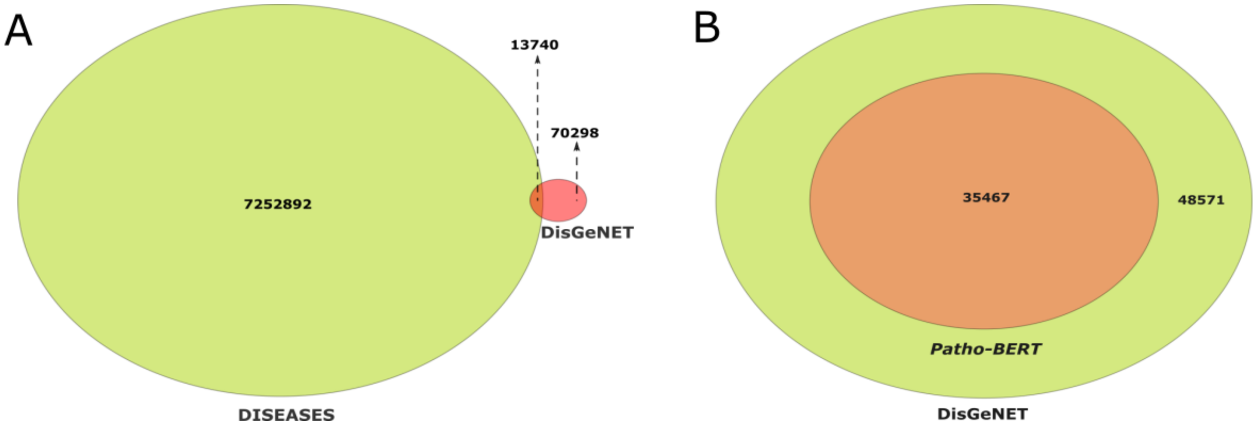

Figure 4.4: DisGeNET and the DISEASES database's gene-disease pairings are intersected. Just ~13,000 out of ~84,000 DisGeNET relationships were found by DISEASES. (B) For almost ~35,000 of the DisGeNET pairs, PathoBERT revealed statistically significant connections. Notably, PathoBERT can generate a disease-gene score for each and every potential connection. It can therefore be utilized to find brand-new associations.

Additionally, it offers empirically derived statistical significance estimates for each *Patho-score*. Figure 4.5 shows the level of gene expression across all organs and the *Patho-score* for APOE. A higher risk of cardiovascular disease is associated with APOE [142]. There may be a gender-specific relationship between the APOE genotype and colon cancer risk, and survival [1].

Additionally, it affects lipid homeostasis. Additionally, the e4 variant of the APOE gene raises a person's risk of developing Alzheimer's disease late in life [2]. Memory loss and other cognitive functions are gradually lost as a result of Alzheimer's disease (AD), which is also characterized by brain atrophy and the development of amyloid plaques [143]. In Figures 4.6 - 4.12, a few additional such relationships are included.

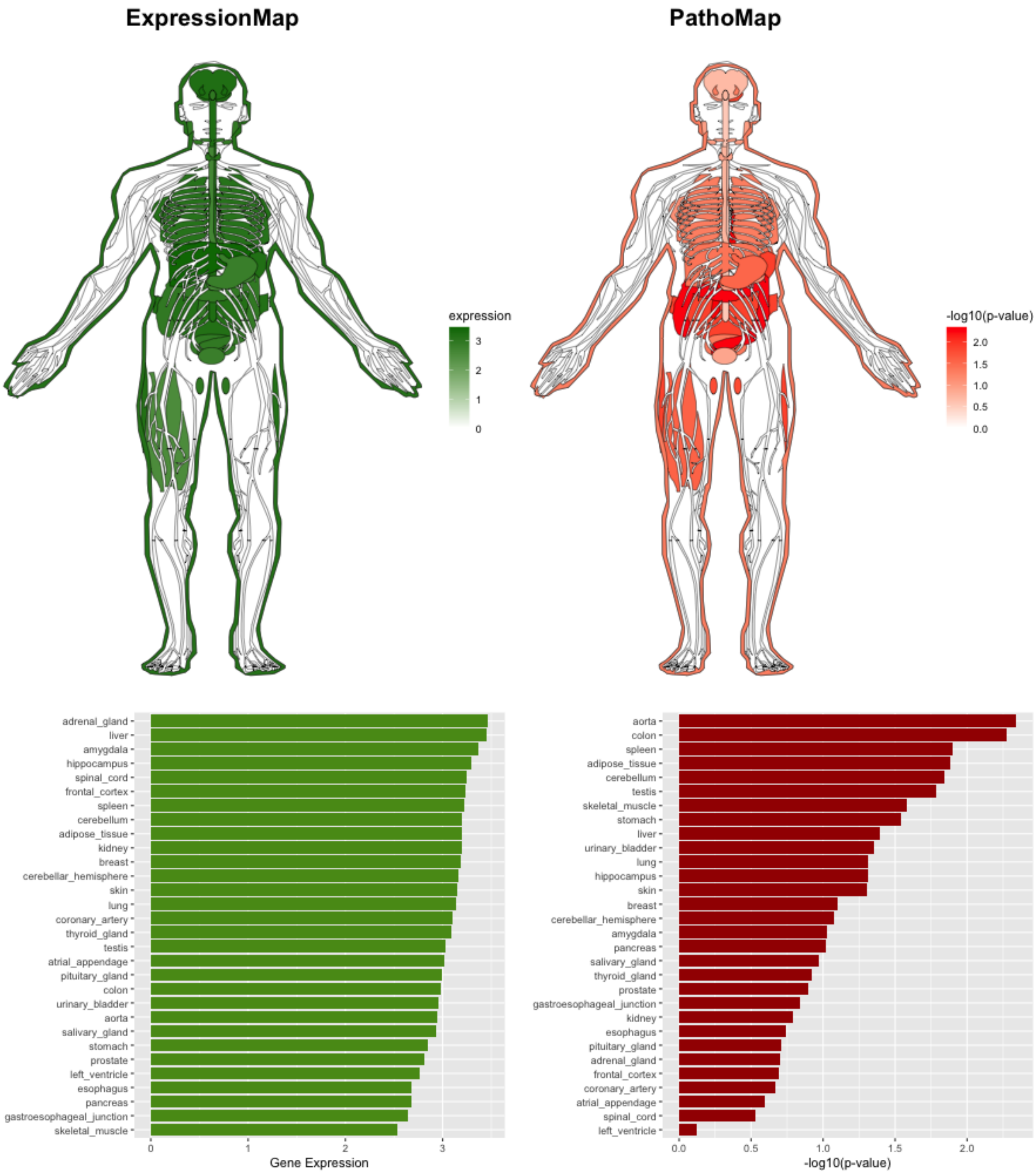

Figure 4.5: The expression and *Patho-scores* of the APOE gene across different human tissues. Tissues have different distributions of statistical significance for *Patho-scores*. There may be a gender-specific relationship between the APOE gene, and the risk of developing colon cancer as well as the prognosis [1]. Also, the presence of the e4 allele of the APOE gene elevates an individual's chance of developing Alzheimer's disease in their later years [2].

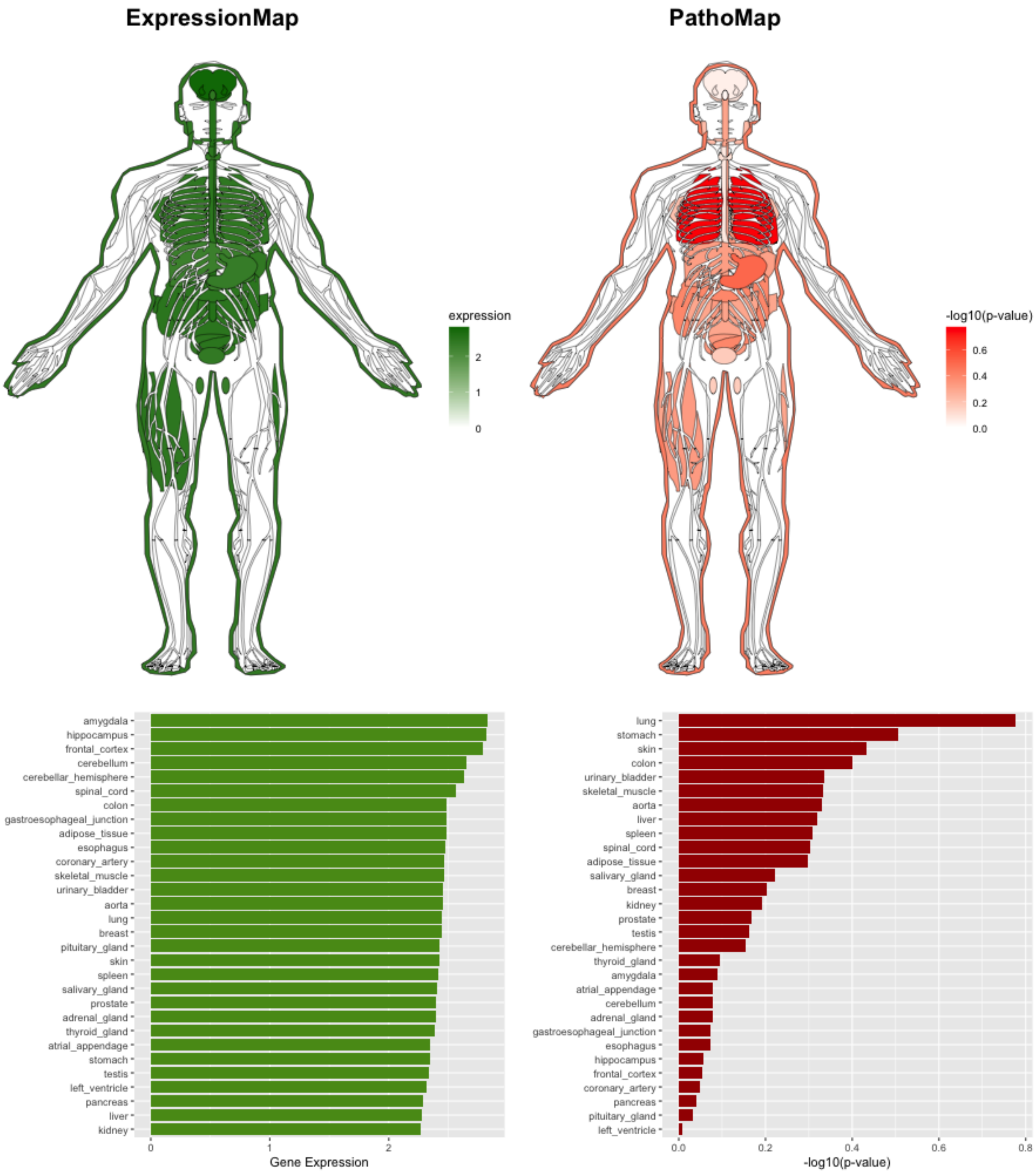

Figure 4.6: The expression and *Patho-scores* of the APC gene across different human tissues. Tissues have different distributions of statistical significance for *Patho-scores*. Germline mutations in the APC gene are the primary cause of familial adenomatous polyposis (FAP) [3]. However, APC gene alterations are proven to have a significant role in cancer pathogenesis and are not just restricted to FAP [4]. Furthermore, APC mutations are a rate-limiting factor in colorectal malignancies [5].

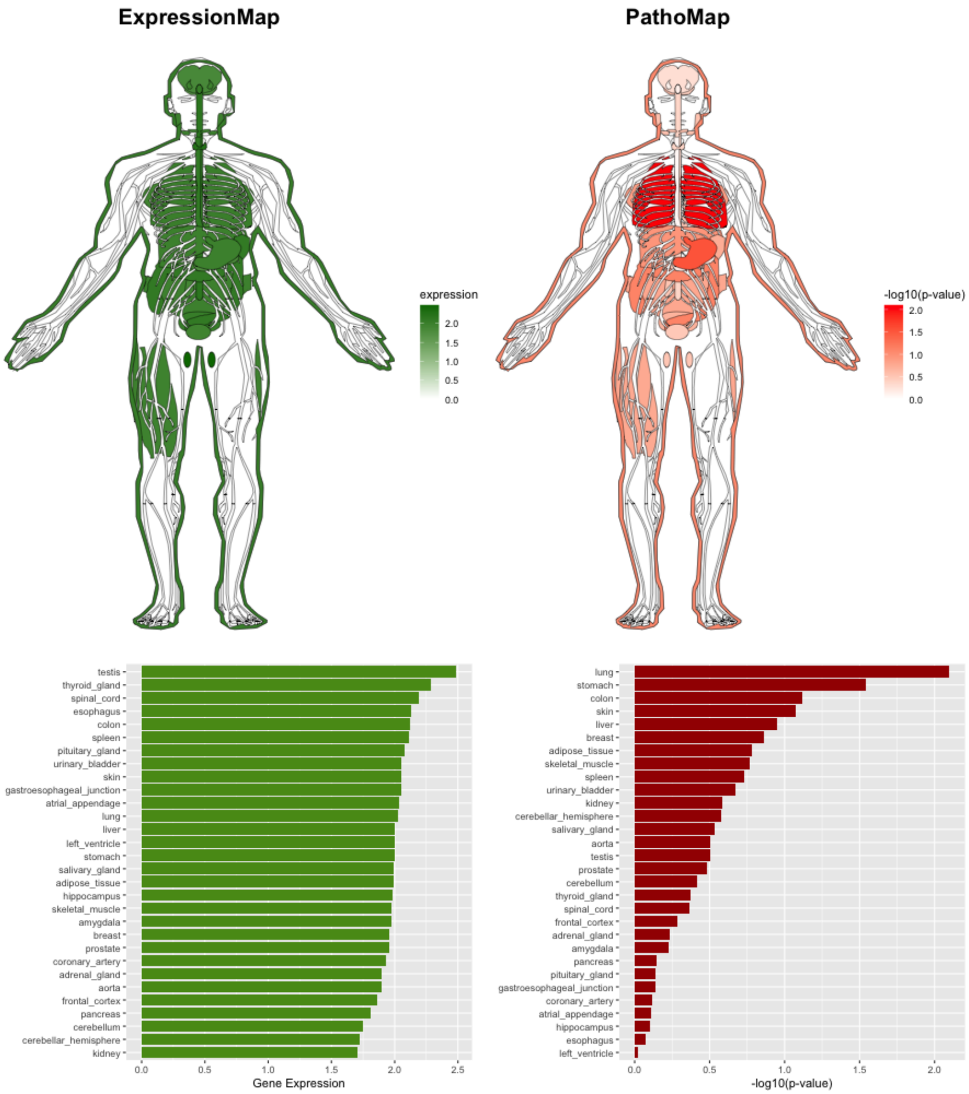

Figure 4.7: The expression and *Patho-scores* of the BRCA1 gene across different human tissues. Tissues have different distributions of statistical significance for *Patho-scores*. The BRCA1 gene is a tumor suppressor. However, changes in this gene increase the susceptibility to a higher risk of developing breast, stomach, colorectal, and other types of cancer [6][7][8]. In non-small-cell lung cancer (NSCLC), BRCA1 overexpression is substantially associated with poor survivability [9].

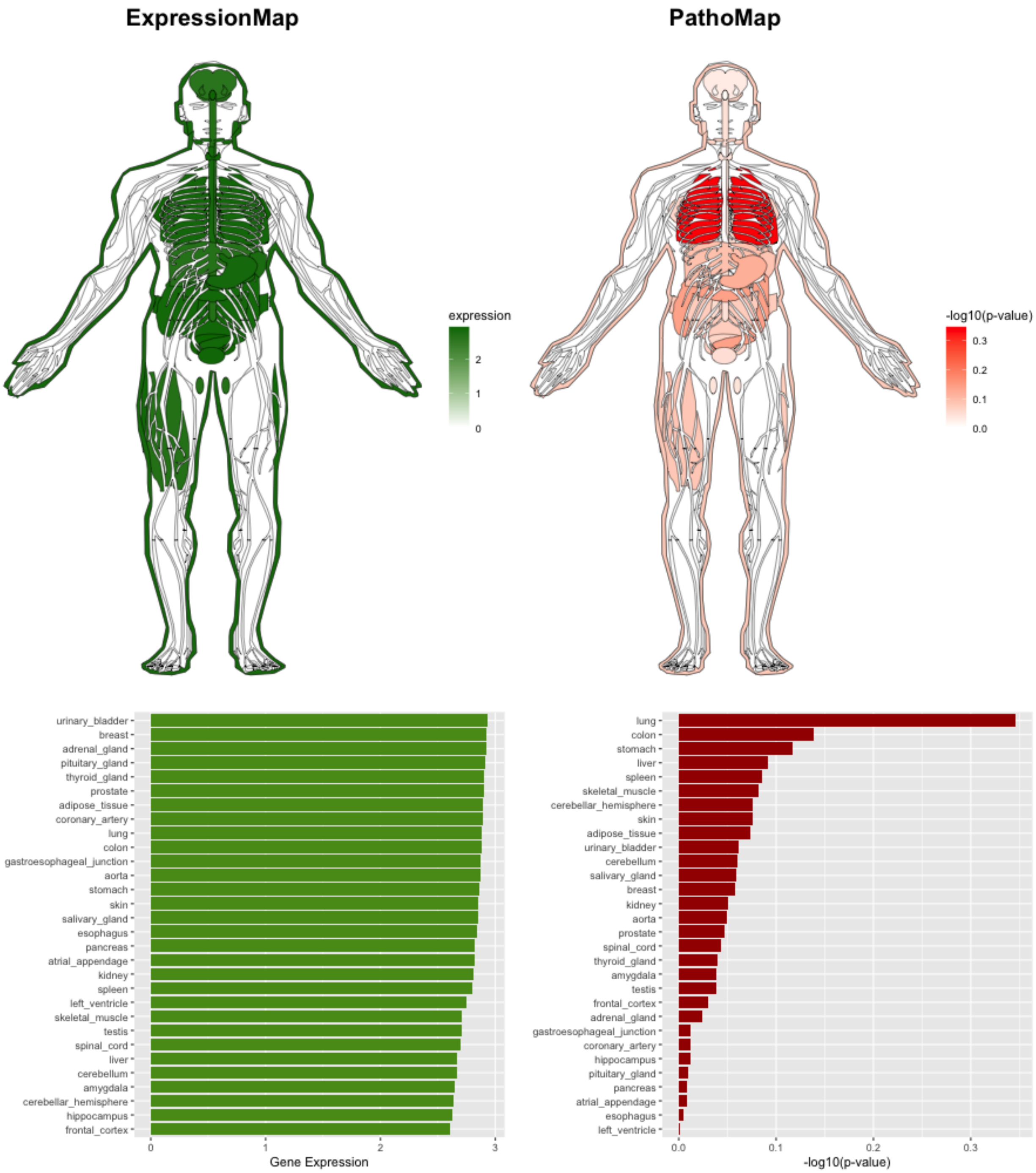

Figure 4.8: The expression and *Patho-scores* of the CDK4 gene across different human tissues. Tissues have different distributions of statistical significance for *Patho-scores*. Lung cancer formation and a poor prognostic have both been linked to overexpression of CDK4 [10].

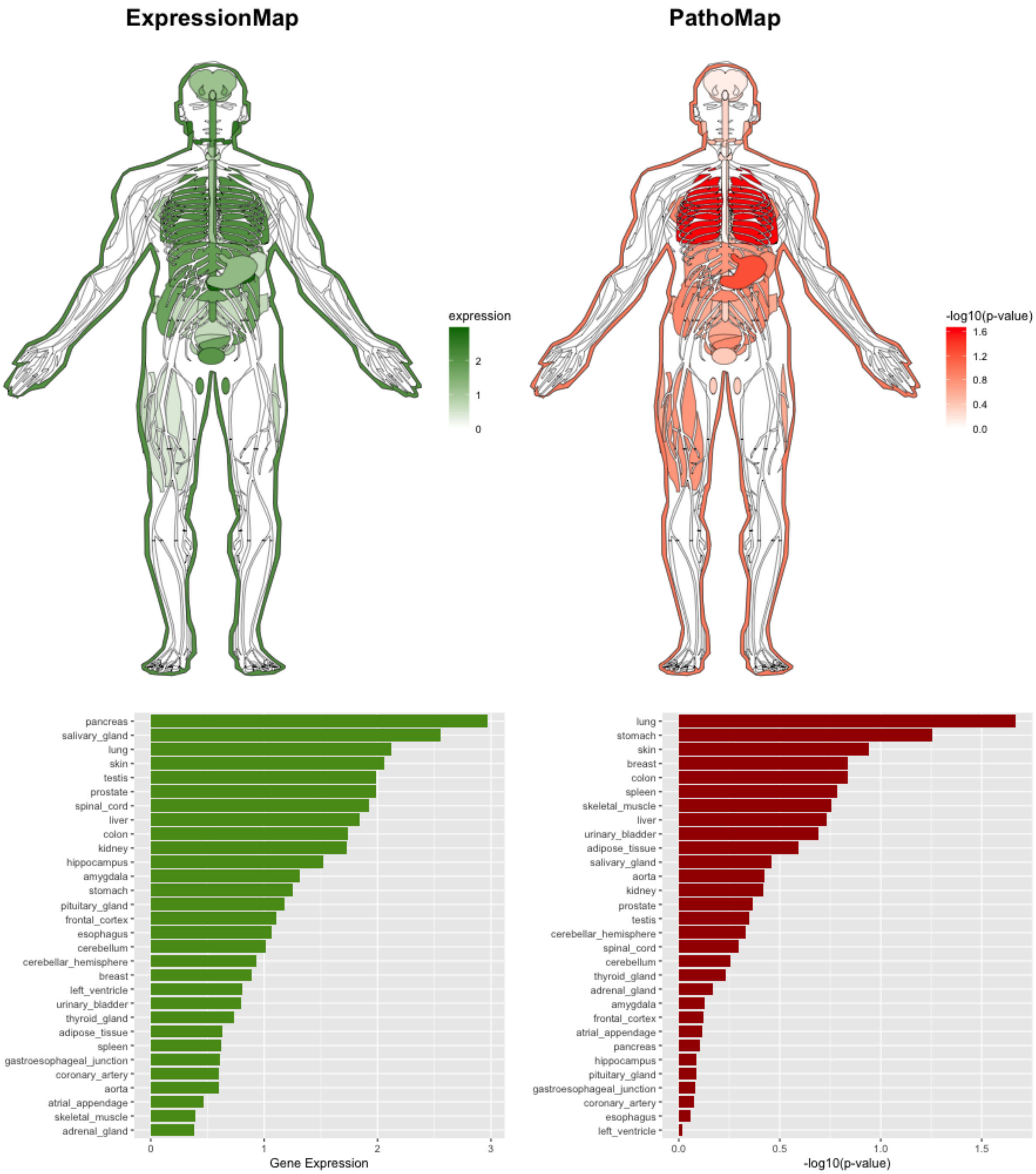

Figure 4.9: The expression and *Patho-scores* of the CFTR gene across different human tissues. Tissues have different distributions of statistical significance for *Patho-scores*. Variations in the CFTR gene are what cause the hereditary fatal illness known as cystic fibrosis (CF) [11]. Ion channels are encoded by the CFTR gene, and disruption or modification in this gene causes a disparity of ions and fluids in tissues including the intestine and airways, among others [12]. Although several tissues are damaged by CF, the lungs have the most detrimental physiological efficacy [13].

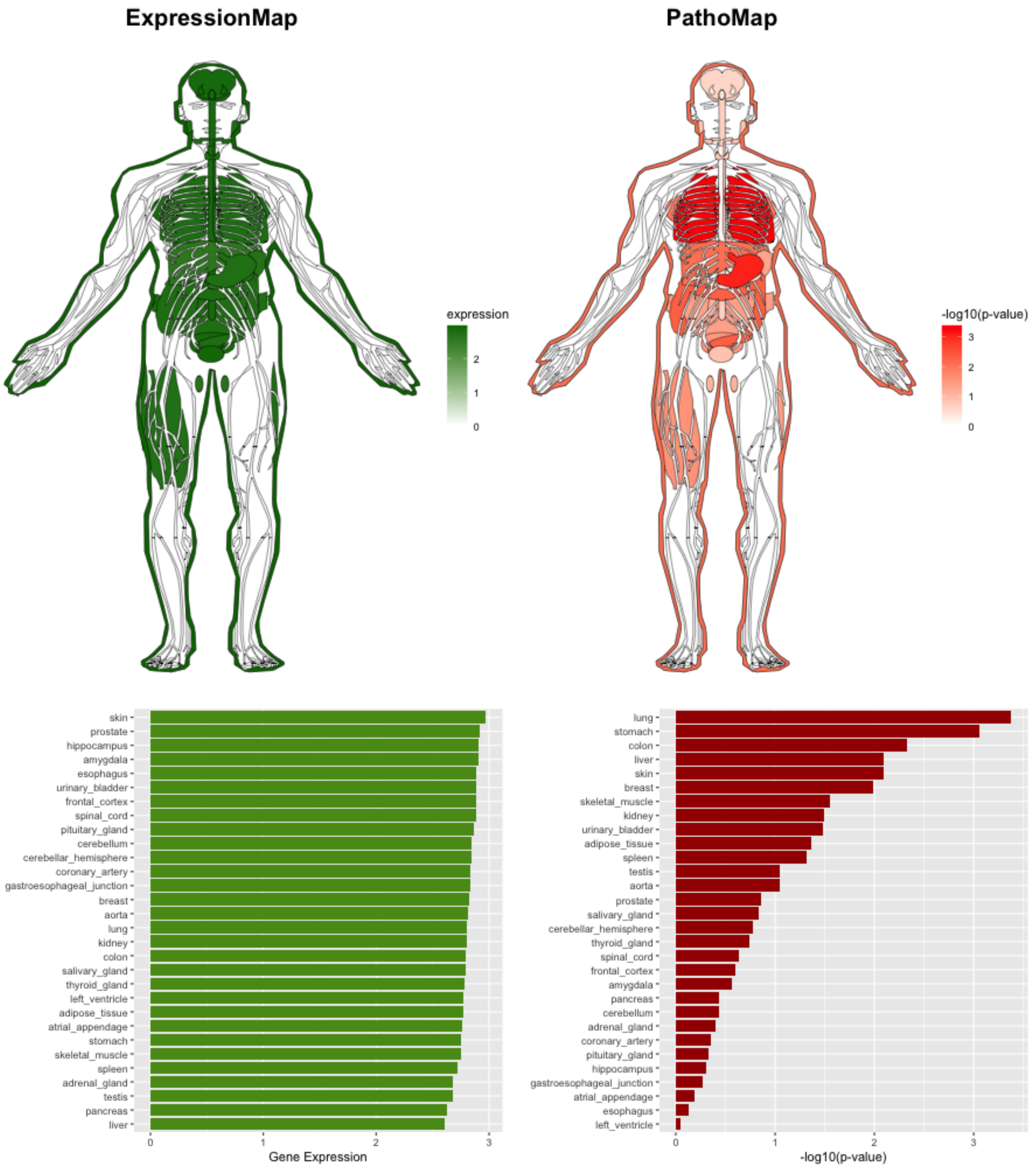

Figure 4.10: The expression and *Patho-scores* of the HRAS gene across different human tissues. Tissues have different distributions of statistical significance for *Patho-scores*. HRAS is linked to disorders including Costello Syndrome [14] and Epidermal Nevus Syndrome, which entails skin-related abnormalities [15]. HRAS gene mutations are prevalent in lung and bladder cancer and may be a possible therapeutic target [16]. Moreover, HRAS amplification is linked to the development and poor prognostic of gastric cancer [17].

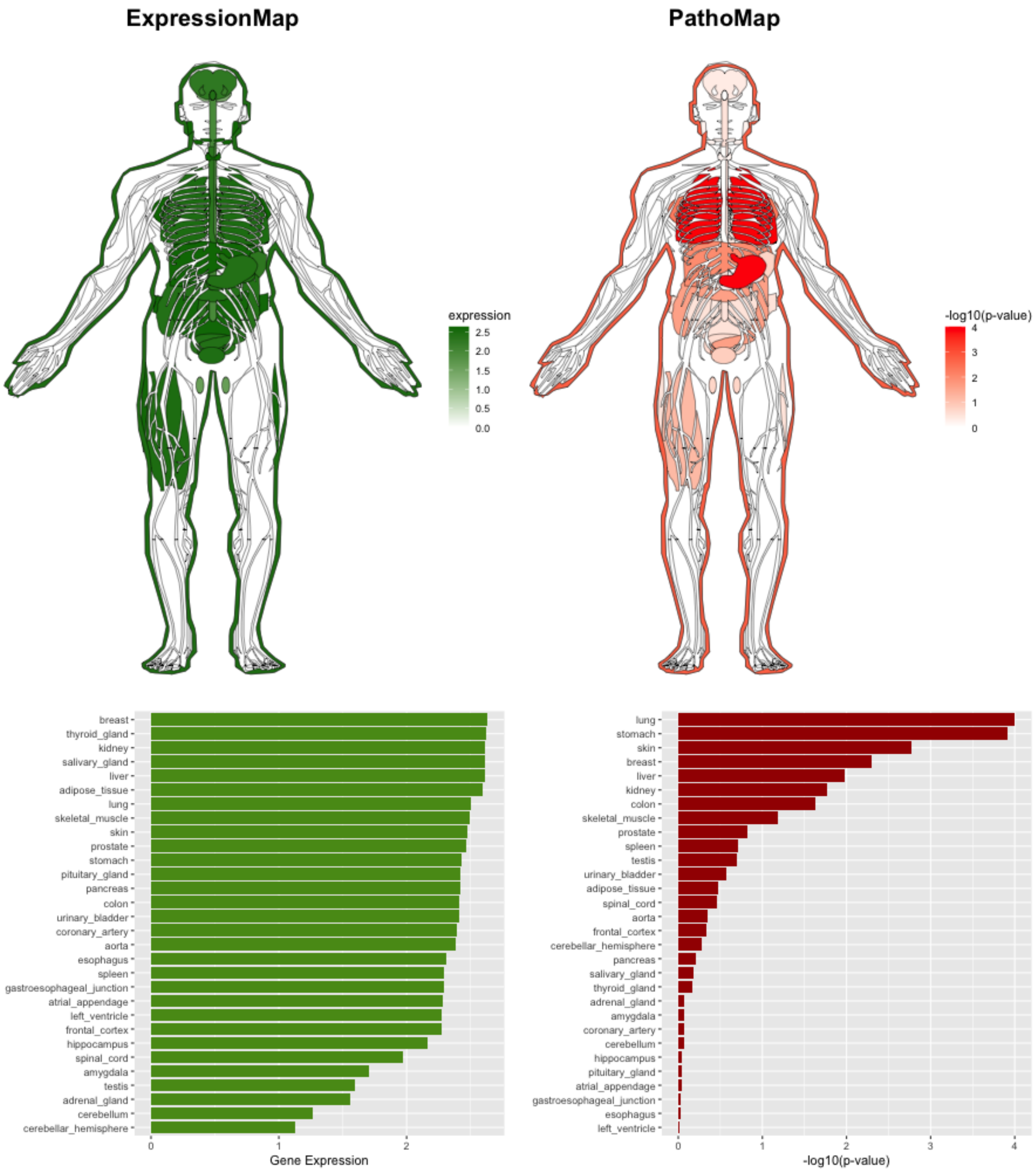

Figure 4.11: The expression and *Patho-scores* of the MET gene across different human tissues. Tissues have different distributions of statistical significance for *Patho-scores*. In numerous solid tumors, MET is either altered or overexpressed. MET activation tends to be a driving force of lung carcinoma and may be a successful therapeutic target for the disease [18]. Moreover, increased MET expression is related to poor survivability of breast cancer survivors [19].

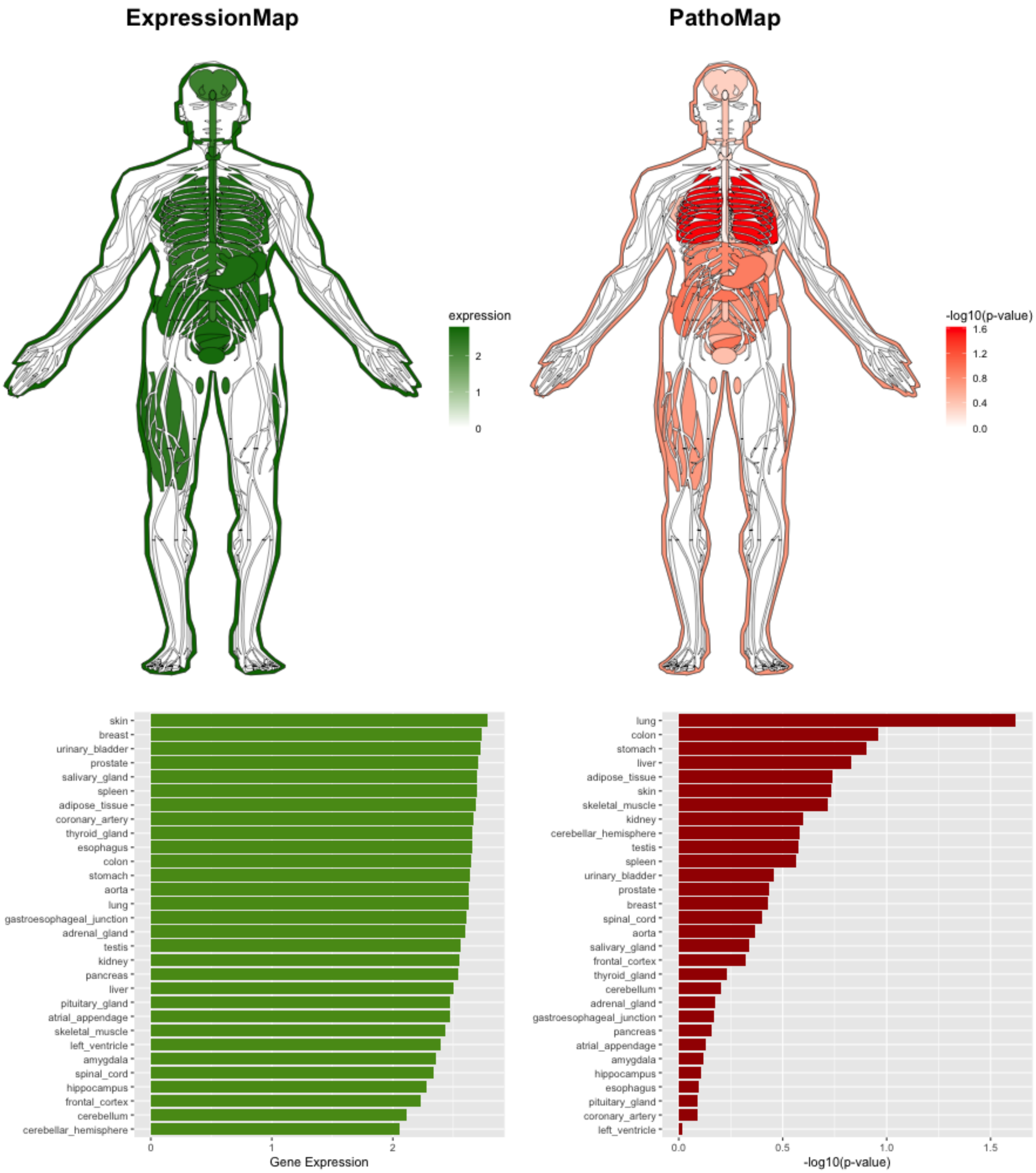

Figure 4.12: The expression and *Patho-scores* of the TP53 gene across different human tissues. Tissues have different distributions of statistical significance for *Patho-scores*. Lung epithelial cell malignant growth is significantly influenced by TP53 gene mutations, which are also associated with a poor prognosis [20]. Modifications to TP53 are also responsible for colon cancer development [21].

We investigated if PathoBERT might be used to identify a gene's cell-type-specific pathogenic roles, as compared to measuring pathological activity specific to an organ. To achieve this, we developed three gene-cell type-disease triplets that include information on autoimmune disease: Sjogren's syndrome, Systemic lupus erythematosus, and rheumatoid arthritis, as well as the impacted cell types and genes implicated in their pathogenesis. T cells, B cells, and macrophages are three immune cells that have been linked to the pathophysiology of Sjogren's disease [144], Systemic lupus erythematosus [145], and rheumatoid arthritis [146], respectively. The Enrichr database provided the details on the genes related to these illnesses [147]. We produced three triplets — SSB - T cells - Sjogren's syndrome; TMC2 - B cells - Systemic lupus erythematosus; ATE1 - Macrophage - Rheumatoid arthritis. We utilized the corresponding *PathoBERT* vectors to illustrate the associations in the PCA space (Figure 4.13). In terms of the relative location of the words across three triplets, we found uniformity in directionalities.

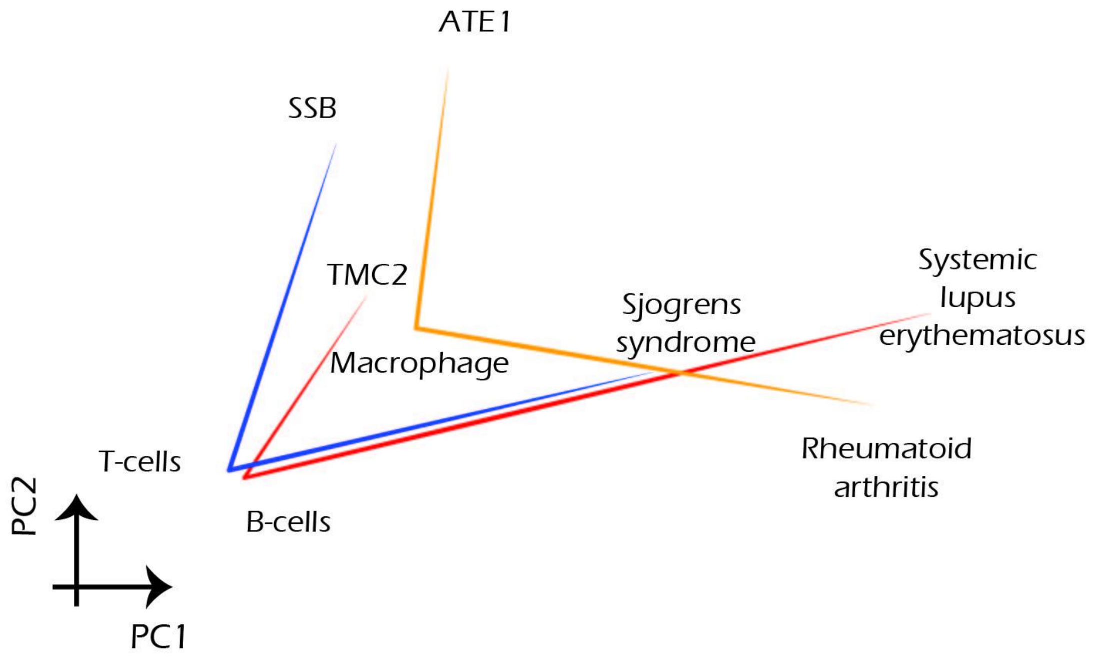

Figure 4.13: Word vectors connected to a well-known gene implicated in illness pathogenesis, the affected cell type, and three triplets of an auto-immune disease.

4.3.5 PathoBERT enables discovery of unknown disease-gene relationships

We anticipated that *PathoBERT* would collect latent knowledge encoded in literature linked to pathology. To test our theory, we created two procedures. First, we sought to determine whether *PathoBERT* can predict disease-gene associations that are not present in our corpus of ~18 million abstracts and have been identified after the end date of abstract archival. The most recent publication in our archive of abstracts was in May 2019. In a paper by Davies et al., published on October 17, 2019, the genomic landscape was molecularly profiled using data from 42 benign and malignant tumors across 13 individuals from four cross-generational families. Reoccurring mutations in the epigenetic modifiers *DNMT3A* and *BCOR* were found in 29% of benign tumors [148]. This was the first study that, as far as we are aware, connected these genes with the *CYLD* cutaneous syndrome (CCS). For the word pairings (CYLD, DNMT3A) and (CYLD, BCOR), we received significant *Patho-scores* with *P-values* of 0.056 and 0.013, respectively.

We used the Random Walk with Restart (RWR) based network diffusion method on a multiplex disease-gene network to show how PathoBERT can be used for the development of novel biomarkers linked with diseases (please refer to Methods section). As seeds for this research, we chose “colorectal cancer” and “*PLA2G2A*” (a gene known to be related to the condition). RWR offers a probability distribution for each disease and gene, giving the seeds' immediate surroundings more weight in the network. The top 15 genes and 15 OMIM disease ids are shown in Figure 4.14A, with relatively high probabilities reflecting frequent access to the seeds. While the majority of the genes in the network are quite well known for their role in colorectal malignancies, we discovered two candidates with scant or no literature support for such a role. These are BAG6 and UCHL3. With *P-value* of 0.019 and 0.040, respectively, the *Patho-scores* between these genes and the “colon” was discovered to be significant (Figure 4.14B). We found that among them, BAG6 upregulation was associated with an abysmal survival rate (Figure 4.14C). In particular, the initial study linking UCHL3 to colorectal cancer was released in 2020 [149]. Although BAG6 has been associated with a higher chance of developing lung cancer [150], no notable research has been found connecting the gene with colorectal cancer. We anticipate that a comprehensive molecular analysis of the gene with a medical standpoint will reveal its mode of action.

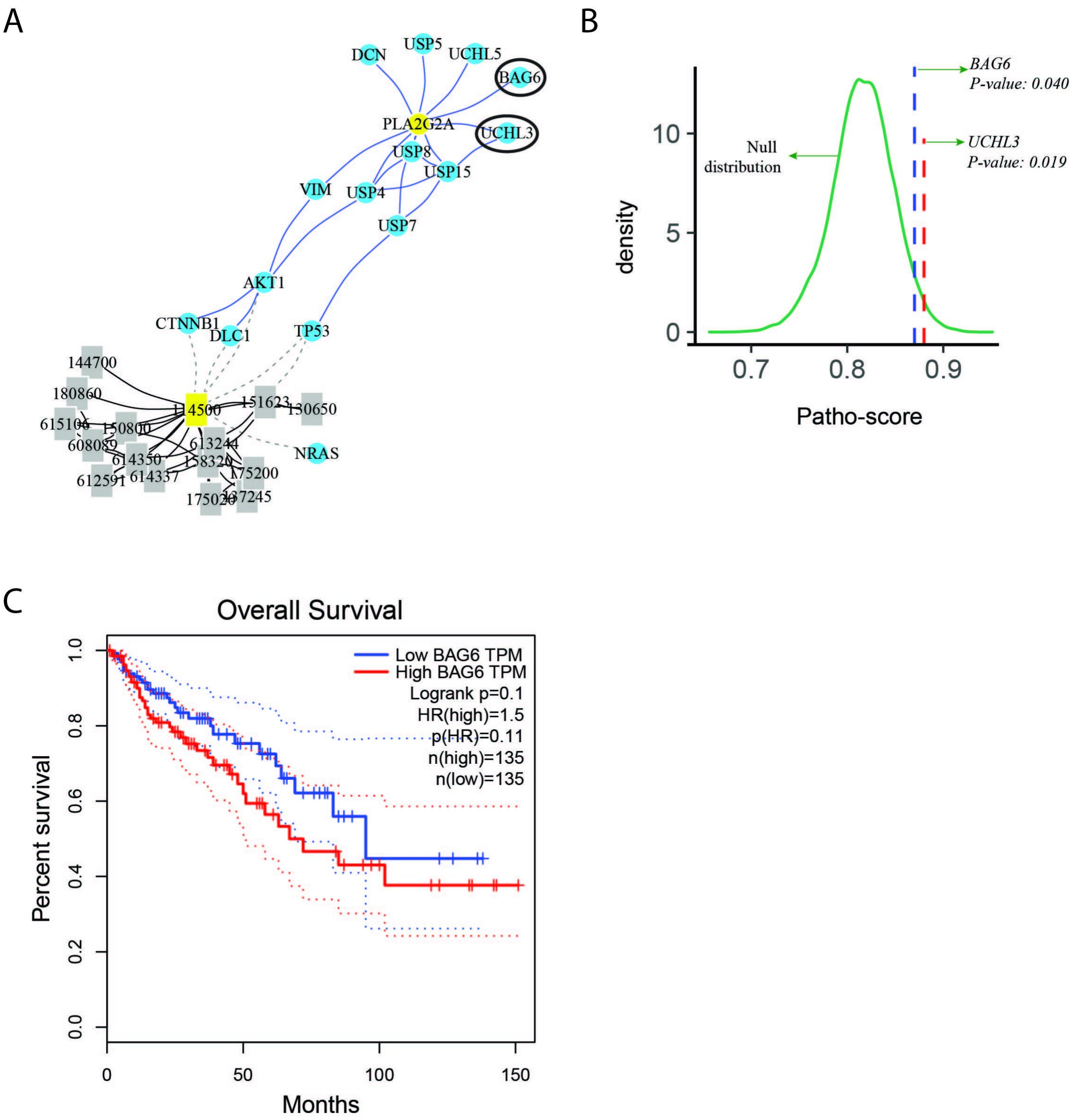

Figure 4.14: (A) In a genome-wide network of proteins, diseases, and nodes with three types of connections - a) disease - disease similarities, b) protein - protein interactions, and c) disease - gene relationships - random walk with restart (RWR) is used. To begin, RWR requires seed nodes, which are planned to be "colorectal cancer" (OMIM: 114500), and PLA2G2A, a gene that has been linked to colorectal cancer. On the graphic, the top 15 genes and disease OMIM IDs are displayed. With the exception of UCHL3 and BAG6, the majority of genes are well known for their relation to colorectal cancer. (B) *P-values* were determined empirically and compared to the null distribution using the permutation test. (C) Based on the degree of BAG6 expression in TCGA colon cancer samples, overall survival was estimated (created using the GEPIA web server).

4.4 Conclusion

Using around 4.5 million abstracts citing disease-gene correlations, we tweaked the BERT base pretrained model. The resulting model, known as *PathoBERT*, was put through a series of inference tasks, including diseases, genes, organs, and cell types. Some of these are examining word analogies integrating various biological concepts - such as disease, genes, and symptoms - as well as gauging the effectiveness of disease-gene relationship prediction. Another is finding novel connections that have either recently emerged in the literature or remain elusive. *PathoBERT* consistently demonstrated significant potential. Essentially, *PathoBERT* and *Pathomap* work together to offer a method for obtaining an impartial continuous representation of disease-causing genes and their activity in particular tissues and cell types. This could hasten focused research into many different diseases. *Patho-scores* and corresponding color intensities show how much a gene contributes to disease in a certain tissue or organ. *Pathomap* does not give specifics regarding a gene's contribution to a certain illness. The web server *DiGSeE* [151] can be used to determine a gene's mode-of-action (such as mutation, regulation, acetylation, or DNA methylation) in a disease.

The present *PathoBERT* model has a significant flaw in that it is not taught to distinguish between different species. As a result, the present scores do not correspond to those of mice, humans, or other primates. After including articles with human-related content, we were left with just about ~1.2 million abstracts, which is insufficient for language modeling tasks. Due to the fact that animal models are normally solely used to imitate human diseases, we are confident that this won't introduce any bothersome variables.

We rigorously compared the abilities of several embedding strategies to infer real relationships between diseases and genes, using humanly curated associations as the gold standard. The results showed that *PathoBERT* was the best method, with BioSentVec coming in second. The remaining strategies, such as Word2vec, mainly fell short. Human cell atlasing is a reality, but a pathological atlas may require a lot of work and resources; therefore, we don't expect it any time soon. We are confident that Pathomap will help the community in reducing the search process for genetic hits until this vision becomes a reality. *PathoBERT* can also be used to choose genes for diagnostic gene panels. We showed that Pathomap could combine hitherto unrecognized relationships between diseases and gene scopes. *PathoBERT* could potentially be used as an additional method to Gene Ontologies for the in-depth analysis of results from high-throughput experiments like RNA-seq or MicroArray. Additionally, it can help prioritize the genes to be chosen for extensive loss of function research. By illuminating the putative function of BAG6 colorectal cancer aetiology and prognosis, we provide a model for such investigations.

Chapter 5

Discussion and Conclusion

5.1 Summary of contribution

On the one hand, where next-generation sequencing (NGS) allows us to distinguish molecular heterogeneity in a tissue's seemingly comparable cells, protocols/technologies like cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq), and RNA expression and protein sequencing assay sequencing (REAP-seq) allows us to map a cell's gene expression with its protein expression and chromatin accessibility. However, this generates a significant amount of single cell data, which eventually results in an increase in the number of publications published each day in peer-reviewed journals.

Despite the fact that single cell datasets give us a high resolution view of the cell states and how they function, they also have a number of serious disadvantages, particularly in terms of dimensionality and sparsity. The data generated using single cell RNA sequencing (scRNA-seq) have a high dimension in terms of the number of genes or features and the number of samples that are processed. This increases the time and complexity needed to process the data. One of the elements that contribute to the sparse nature of single cell data is the biological noise, which is introduced during the generation stage together with the inherent technical noise. It is possible that a gene's expression will not be detected even though it is actively expressed in the cell because of the incredibly low amount of RNA that is found in a single cell. Dropout occurrences arise as a result of this, which also adds to the data's sparseness. Eliminating meaningless and redundant features that might hide an important group of genes in irrelevant or noisy data is critical. This motivated us to work on a feature selection method, SelfE, an $l_{2,0}$-minimization technique that identifies the optimum subset of feature vectors while maintaining the observed subspace structures in the data.

Finding and displaying transcriptionally comparable cells for proper investigation of the cellular diversity provided by single cell transcriptomics is another significant issue that must be solved to significantly improve the downstream analysis. Our next work expands on the deep dictionary methodology for clustering cells that are similar in nature. In this cluster-aware deep dictionary learning framework, we incorporated K-means clustering and sparse subspace clustering as the loss functions to attain meaningful cell clades.

Numerous publications are published each day in scholarly journals as a direct result of the high throughput single cell technologies that generate massive amounts of single data. The sheer volume of manuscripts makes it challenging to assemble and make sense of relevant data. Therefore, in our next study, we trained a text mining model, PathoBERT, which is trained on ~18 million PubMed abstracts to learn the knowledge residing in the specific biomedical literature. The intention was to create a comprehensive resource for understanding the tissue-specific pathogenic actions of genes. It permits querying a gene to observe the impact of tissue-specific pathogenic actions of a particular gene on a human body layout. Parallel to the Human Cell Atlas project, which is still integrating the expression of genes in normal healthy tissues, Pathomap gives insights into tissue-specific pathogenic gene activity.

Our studies collectively use feature engineering techniques for the representation of biological entities in low-dimensional space, which is briefly summarized in the upcoming sections.

5.2 SelfE: Gene selection via self expression for single cell data

Numerous innovative techniques for reducing the number of features have been put forth in the literature. It can be difficult for physicians to assess the limited feature space that a feature extraction technique offers. By creating a method for selecting a minimal feature subset that includes information from the entire feature space, this study contributes to the feature selection domain. Selecting genes with variability is the general guiding idea. However, this approach has a limitation in that it is susceptible to data noise. There is a chance that this will result in false alarms for expression variability. Because there could be numerous additional confounders, accounting for average expression may not be sufficient in many circumstances. It should be noted that genes that are constantly expressed, such as housekeeping genes, can be described by just one or a small number of genes in the condensed feature set produced by SelfE, based on how much value we give to the sparsity term. As a result, it would be unusual for SelfE to devote a significant number of genes to understanding genes expressed steadily.

With regard to the validation metrics - clustering accuracy, distance preservation, and cell visualization - this research made a methodological contribution to feature selection that surpasses entire gene sets and conventional methodologies. SelfE consistently returns the original labels with a wide range of genes picked, making it the most trustworthy of the examined techniques. It should be emphasized that most methods work just as well if more than 200 genes are selected. The idea of the union of subspaces serves as the foundation for the provided feature selection technique. It assumes that the samples are organized into different subspaces and extracts from the data the least squares-efficient representational basis for each subspace. It is fundamentally distinct from current widely used methods like the coefficient of variation, PCA loadings, Gini coefficient, and Fano factor since it is based on linear algebraic approaches as opposed to prior statistical models.

SelfE is predicted to eclipse other techniques as the method of choice for locating meaningful features in single cell data. In the future, we'll try a number of non-linear kernels to enhance its efficiency further.

5.3 Cluster aware deep dictionary learning for single cell analysis

This work proposes a deep dictionary learning-based clustering method. The original single cell data is transformed into a low-dimensional embedding (where cells are represented in rows and genes are represented in columns), which is then inputted to a clustering algorithm. A low-dimensional embedding that is trained to produce intuitively clustered output is used to represent each sample. In order to evaluate the performance of the proposed method, we compared it to state-of-the-art deep learning methods (stacked autoencoder and deep belief network) and specialized single-cell RNA clustering techniques (GiniClust, seurat and SC3). Two clustering accuracy measures, namely, adjusted rand index (ARI) and normalized mutual information (NMI) was used to assess the performance of different clustering algorithms.

The current strategy is greedy and hence suboptimal because there is no interaction between the deep and shallow layers. In future study, the cutting edge optimization tools can be used to address the same.

5.4 Literature mining discerns latent disease gene relationships

To comprehend the immense amount of data that researchers have amassed, we amended the BERT base pretrained model using about 4.5 million abstracts citing disease-gene connections. The final model, called *PathoBERT*, was subjected to various inference tasks, including diseases, genes, organs, and cell types. Some of these include testing disease-gene association prediction accuracy and looking at word analogies integrating numerous biological concepts, such as disease, genes, and symptoms. Another is stumbling onto fresh links that have either recently appeared in the literature or continue to elude us. *PathoBERT* constantly showed tremendous promise. In essence, *PathoBERT* and *Pathomap* provide a means for acquiring a continuous, unbiased depiction of disease-causing genes and their activity in specific tissues and cell types. This might speed up the targeted study into a variety of ailments. *Patho-scores* and associated color intensities reflect the degree to which a gene causes disease in a particular tissue or organ.

Pathomap does not provide details about how a gene may have contributed to a certain disease. To ascertain a gene's mode of action in a disease (such as mutation, regulation, acetylation, or DNA methylation), one can use the *DiGSeE* web server.

The current *PathoBERT* model has a serious shortcoming: it is not taught how to identify one species from another. Because of this, the current scores do not reflect those of mice, humans, or other primates. We only have ~1.2 million abstracts after including papers with human-related content, which is inadequate for text processing tasks. We are convinced that this won't create any problematic factors because animal models are typically only used to mimic human diseases.

Using human-curated associations as the reference standard, we thoroughly examined the capacities of several embedding techniques to infer true correlations between illnesses and genes. The outcomes demonstrated that *PathoBERT* was the most effective technique, with BioSentVec placing second. The remaining approaches, like Word2vec, were mostly ineffective. Human cell atlasing is an actuality, but since creating a pathological atlas might take significant time and effort, we don't see it happening very soon. Until this ambition becomes an actuality, we are convinced that Pathomap will assist the community in speeding up the process of looking for genetic hits. The selection of genes for diagnostic genetic panels can also be done using *PathoBERT*. We demonstrated how Pathomap might incorporate previously unknown connections between diseases and gene sets. For the comprehensive study of outcomes from high-throughput studies like RNA-seq or MicroArray, *PathoBERT* may be utilized in addition to Gene Ontologies. It can also aid in prioritizing the genes to be selected for in-depth loss of function studies. We offer a model for such investigations by outlining the potential role of BAG6 colorectal cancer etiology and prognosis.

5.5 Future work

The recent technologies such as REAP-seq (RNA expression and protein sequencing assay sequencing) [152] and CITE-seq (cellular indexing of transcriptomes and epitopes by sequencing) [153] allows us to profile both the gene expression and protein expression of a cell simultaneously. The amount of mRNA present in a single cell during the reverse transcription process is very low due to which it might not get detected. This leads to zero expression for the particular gene in a single cell, eventually making the single cell gene expression data sparse. Apart from this, various technical and biological variations lead to the sparsity of single cell data. On the other hand, the protein expression data is incomplete as fewer proteins are profiled for a single cell. In this work, we propose a collaborative matrix completion framework that performs matrix completion on both transcriptomic and proteomic data, using cell information for the proteomic and transcriptomic counterparts. The proposed framework will make use of graph regularized nuclear norm minimization for simultaneous imputation of the gene expression and protein expression data. For identifying the missing values, we will employ graph laplacian from the transcriptomics and proteomics data. We will use both the graph laplacian to co-complete gene and protein data and compare it with a scenario when just graph laplacian obtained using gene expression data is used to impute transcriptomics data, and graph laplacian obtained using protein expression data is used to impute proteomics data. The imputation results obtained from different scenarios can be evaluated on the basis of different metrics like clustering accuracy (adjusted rand index and normalized mutual information) and two dimensional visualization.

References

1. Watson MA, Gay L, Stebbings WS, Speakman CT, Bingham SA, Loktionov A. Apolipoprotein E gene polymorphism and colorectal cancer: gender-specific modulation of risk and prognosis. Clin Sci . 2003;104. doi:10.1042/CS20020329

2. Liu CC, Liu CC, Kanekiyo T, Xu H, Bu G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nat Rev Neurol. 2013;9. doi:10.1038/nrneurol.2012.263

3. Leoz ML, Carballal S, Moreira L, Ocaña T, Balaguer F. The genetic basis of familial adenomatous polyposis and its implications for clinical practice and risk management. Appl Clin Genet. 2015;8. doi:10.2147/TACG.S51484

4. Fang DC, Luo YH, Yang SM, Li XA, Ling XL, Fang L. Mutation analysis of APC gene in gastric cancer with microsatellite instability. World J Gastroenterol. 2002;8. doi:10.3748/wjg.v8.i5.787

5. Fodde R. The APC gene in colorectal cancer. Eur J Cancer. 2002;38. doi:10.1016/s0959-8049(02)00040-0

6. Godet I, Gilkes DM. BRCA1 and BRCA2 mutations and treatment strategies for breast cancer. Integrative cancer science and therapeutics. 2017;4. doi:10.15761/ICST.1000228

7. Levine DA, Argenta PA, Yee CJ, Marshall DS, Olvera N, Bogomolniy F, et al. Fallopian Tube and Primary Peritoneal Carcinomas Associated With BRCA Mutations. J Clin Oncol. 2016 [cited 7 Oct 2022]. doi:10.1200/JCO.2003.04.131

8. Website. Available: https://onlinelibrary.wiley.com/doi/10.1111/cge.12497

9. Rosell R, Skrzypski M, Jassem E, Taron M, Bartolucci R, Sanchez JJ, et al. BRCA1: A Novel Prognostic Factor in Resected Non-Small-Cell Lung Cancer. PLoS One. 2007;2. doi:10.1371/journal.pone.0001129

10. Wu A, Wu B, Guo J, Luo W, Wu D, Yang H, et al. Elevated expression of CDK4 in lung cancer. J Transl Med. 2011;9: 1–9.

11. Mall MA, Hartl D. CFTR: cystic fibrosis and beyond. Eur Respir J. 2014;44. doi:10.1183/09031936.00228013

12. Lopes-Pacheco M. CFTR Modulators: The Changing Face of Cystic Fibrosis in the Era of Precision Medicine. Front Pharmacol. 2020;10. doi:10.3389/fphar.2019.01662

13. Fraser-Pitt D, O'Neil D. Cystic fibrosis – a multiorgan protein misfolding disease. Future Science OA. 2015;1. doi:10.4155/fso.15.57

14. Estep AL, Tidyman WE, Teitell MA, Cotter PD, Rauen KA. HRAS mutations in Costello syndrome: detection of constitutional activating mutations in codon 12 and 13 and loss of wild-type allele in malignancy. Am J Med Genet A. 2006;140. doi:10.1002/ajmg.a.31078

15. Avitan-Hersh E, Tatur S, Indelman M, Gepstein V, Shreter R, Hershkovitz D, et al. Postzygotic HRAS mutation causing both keratinocytic epidermal nevus and thymoma and associated with bone dysplasia and hypophosphatemia due to elevated FGF23. J Clin Endocrinol Metab. 2014;99. doi:10.1210/jc.2013-2813

16. Kiessling MK, Curioni-Fontecedro A, Samaras P, Atrott K, Cosin-Roger J, Lang S, et al. Mutant HRAS as novel target for MEK and mTOR inhibitors. Oncotarget. 2015;6: 42183.

17. Wu XY, Liu WT, Wu ZF, Chen C, Liu JY, Wu GN, et al. Identification of HRAS as cancer-promoting gene in gastric carcinoma cell aggressiveness. Am J Cancer Res. 2016;6: 1935.

18. Drilon A, Cappuzzo F, Ou SI, Camidge DR. Targeting MET in Lung Cancer: Will Expectations Finally Be MET? J Thorac Oncol. 2017;12. doi:10.1016/j.jtho.2016.10.014

19. de Melo Gagliato D, Jardim DL, Falchook G, Tang C, Zinner R, Wheler JJ, et al. Analysis of MET genetic aberrations in patients with breast cancer at MD Anderson Phase I unit. Clin Breast Cancer. 2014;14. doi:10.1016/j.clbc.2014.06.001

20. Mogi A, Kuwano H. TP53 mutations in nonsmall cell lung cancer. J Biomed Biotechnol. 2011;2011. doi:10.1155/2011/583929

21. Williams DS, Mouradov D, Browne C, Palmieri M, Elliott MJ, Nightingale R, et al. Overexpression of TP53 protein is associated with the lack of adjuvant chemotherapy benefit in patients with stage III colorectal cancer. Mod Pathol. 2019;33: 483–495.

22. Watson JD, Crick FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid (Reprinted from Nature, April 25, 1953). Nature. 1969. pp. 470–471. doi:10.1038/224470a0

23. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10: 57–63.

24. Wang Y, Navin NE. Advances and Applications of Single-Cell Sequencing Technologies. Molecular Cell. 2015. pp. 598–609. doi:10.1016/j.molcel.2015.05.005

25. Auer PL, Doerge RW. Statistical Design and Analysis of RNA Sequencing Data. Genetics. 2010. pp. 405–416. doi:10.1534/genetics.110.114983

26. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods. 2014;11: 637–640.

27. Liu S, Trapnell C. Single-cell transcriptome sequencing: recent advances and remaining challenges. F1000Res. 2016;5. doi:10.12688/f1000research.7223.1

28. Adam M, Potter AS, Steven Potter S. Psychrophilic proteases dramatically reduce single cell RNA-seq artifacts: A molecular atlas of kidney development. Development. 2017. doi:10.1242/dev.151142

29. Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics. 2018;19: 562–578.

30. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16: 241.

31. van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell. 2018;174: 716–729.e27.

32. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, et al. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods. 2018;15: 539–542.

33. Wang J, Agarwal D, Huang M, Hu G, Zhou Z, Ye C, et al. Data Denoising with transfer learning in single-cell transcriptomics. doi:10.1101/457879

34. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun. 2018;9: 997.

35. Andrews TS, Hemberg M. False signals induced by single-cell imputation. F1000Res. 2018;7: 1740.

36. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161: 1187–1201.

37. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015;347: 1138–1142.

38. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344: 1396–1401.

39. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature. 2014;509: 371–375.

40. Usoskin D, Furlan A, Islam S, Abdo H, Lönnerberg P, Lou D, et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat Neurosci. 2015;18: 145–153.

41. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 2015;33: 155–160.

42. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32: 381–386.

43. Ponting CP. The Human Cell Atlas: making "cell space" for disease. Disease Models & Mechanisms. 2019. doi:10.1242/dmm.037622

44. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017;357: 661–667.

45. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res. 2016;5: 2122.

46. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat Methods. 2017;14: 565–571.

47. Ntranos V, Kamath GM, Zhang JM, Pachter L, Tse DN. Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. Genome Biol. 2016;17: 112.

48. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 2011;21: 1543–1551.

49. Zhang JM, Fan J, Fan HC, Rosenfeld D, Tse DN. An interpretable framework for clustering single-cell RNA-Seq datasets. BMC Bioinformatics. 2018;19: 93.

50. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. PLoS Comput Biol. 2015;11: e1004575.

51. Prabhakaran S, Azizi E, Carr A, Pe'er D. Dirichlet Process Mixture Model for Correcting Technical Variation in Single-Cell Gene Expression Data. JMLR Workshop Conf Proc. 2016;48: 1070–1079.

52. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015;525: 251–255.

53. Zhang H, Lee CAA, Li Z, Garbe JR, Eide CR, Petegrosso R, et al. A multitask clustering approach for single-cell RNA-seq analysis in Recessive Dystrophic Epidermolysis Bullosa. PLoS Comput Biol. 2018;14: e1006053.

54. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14: 483–486.

55. Žurauskienė J, Yau C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. BMC Bioinformatics. 2016;17: 140.

56. Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. Fast interpolation-based

t-SNE for improved visualization of single-cell RNA-seq data. Nature Methods. 2019. pp. 243–245. doi:10.1038/s41592-018-0308-4

57. Lin C, Jain S, Kim H, Bar-Joseph Z. Using neural networks for reducing the dimensions of single-cell RNA-Seq data. Nucleic Acids Res. 2017;45: e156.

58. Ding J, Condon A, Shah SP. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat Commun. 2018;9: 2002.

59. Sinha D, Kumar A, Kumar H, Bandyopadhyay S, Sengupta D. dropClust: efficient clustering of ultra-large scRNA-seq data. Nucleic Acids Res. 2018;46: e36.

60. Jiang L, Chen H, Pinello L, Yuan G-C. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. 2016;17: 144.

61. Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, et al. A survey of human brain transcriptome diversity at the single cell level. Proc Natl Acad Sci U S A. 2015;112: 7285–7290.

62. Tools for Single Cell Genomics. [cited 7 Oct 2022]. Available: https://satijalab.org/seurat/

63. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36: 1234–1240.

64. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 2013 [cited 7 Oct 2022]. doi:10.48550/arXiv.1301.3781

65. Altszyler E, Sigman M, Ribeiro S, Slezak DF. Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. 2016 [cited 17 Oct 2022]. doi:10.48550/arXiv.1610.01520

66. Comparative study of word embedding methods in topic segmentation. Procedia Comput Sci. 2017;112: 340–349.

67. Elekes Á, Englhardt A, Schäler M, Böhm K. Toward meaningful notions of similarity in NLP embedding models. International Journal on Digital Libraries. 2018;21: 109–128.

68. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018 [cited 7 Oct 2022]. doi:10.48550/arXiv.1810.04805

69. [No title]. [cited 17 Oct 2022]. Available: http://josecamachocollados.com/book_embNLP_draft.pdf

70. Taylor WL. Cloze procedure : A new tool for measuring readability. Journal Q. 1953;30: 415–433.

71. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data. 2019;6: 52.

72. Rosenthal G, Váša F, Griffa A, Hagmann P, Amico E, Goñi J, et al. Mapping higher-order relations between brain structure and function with embedded vector representations of connectomes. Nat Commun. 2018;9: 2178.

73. Srinivasan P, Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics. 2004;20 Suppl 1. doi:10.1093/bioinformatics/bth914

74. Ahlers CB, Hristovski D, Kilicoglu H, Rindflesch TC. Using the Literature-Based Discovery Paradigm to Investigate Drug Mechanisms. AMIA Annu Symp Proc. 2007;2007: 6.

75. Wan F, Zhu Y, Hu H, Dai A, Cai X, Chen L, et al. DeepCPI: A Deep Learning-based Framework for Large-scale in silico Drug Screening. Genomics Proteomics Bioinformatics. 2019;17: 478–495.

76. Agarwal P, Searls DB. Literature mining in support of drug discovery. Brief Bioinform. 2008;9: 479–492.

77. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472: 90–94.

78. Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, et al. Author Correction: Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet. 2018;50: 1754.

79. Srivastava D, Iyer A, Kumar V, Sengupta D. CellAtlasSearch: a scalable search engine for single cells. Nucleic Acids Res. 2018;46: W141–W147.

80. Tirosh I, Izar B, Prakadan SM, Wadsworth MH 2nd, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352: 189–196.

81. Zhou B, Jin W. Visualization of Single Cell RNA-Seq Data Using t-SNE in R. Methods Mol Biol. 2020;2117: 159–167.

82. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. Publisher Correction: A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2019;10: 646.

83. Yu L, Liu H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Proceedings, Twentieth International Conference on Machine Learning. 2003. pp. 856–863.

84. Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence. 1997. pp. 273–324. doi:10.1016/s0004-3702(97)00043-x

85. Lal TN, Chapelle O, Weston J, Elisseeff A. Embedded Methods. Feature Extraction. 2006; 137–165.

86. Ni W. A Review and Comparative Study on Univariate Feature Selection Techniques. 2012.

87. Blainey PC, Quake SR. Dissecting genomic diversity, one cell at a time. Nat Methods. 2013;11: 19–21.

88. Tropp JA, Gilbert AC, Strauss MJ. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. Signal Processing. 2006;86: 572–588.

89. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8: 14049.

90. McDaid AF, Greene D, Hurley N. Normalized Mutual Information to evaluate overlapping community finding algorithms. 2011 [cited 7 Oct 2022]. doi:10.48550/arXiv.1110.2515

91. Santos JM, Embrechts M. On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. Artificial Neural Networks – ICANN 2009. 2009; 175–184.

92. Jindal A, Gupta P, Jayadeva, Sengupta D. Discovery of rare cells from voluminous single cell expression data. Nat Commun. 2018;9: 4719.

93. Sengupta D, Rayan NA, Lim M, Lim B, Prabhakar S. Fast, scalable and accurate differential expression analysis for single cells. 2016 [cited 7 Oct 2022]. doi:10.1101/049734

94. Lance Parsons Arizona State University, Tempe, AZ, Ehtesham Haque Arizona State University, Tempe, AZ, Huan Liu Arizona State University, Tempe, AZ. Subspace clustering for high dimensional data. ACM SIGKDD Explorations Newsletter. 2004 [cited 7 Oct 2022]. doi:10.1145/1007730.1007731

95. A review of clustering techniques and developments. Neurocomputing. 2017;267: 664–681.

96. Data clustering: 50 years beyond K-means. Pattern Recognit Lett. 2010;31: 651–666.

97. [No title]. [cited 7 Oct 2022]. Available: https://www.cs.utexas.edu/users/inderjit/public_papers/kdd_spectral_kernelkmeans.pdf

98. [No title]. [cited 10 Jun 2023]. Available: https://www-ai.cs.tu-dortmund.de/LEHRE/FACHPROJEKT/SS14/Papers/Clustering/SpectralClustering.pdf

99. Subspace Clustering. [cited 7 Oct 2022]. Available: https://ieeexplore.ieee.org/document/5714408

100. Sparse subspace clustering. [cited 7 Oct 2022]. Available:

https://ieeexplore.ieee.org/document/5206547

101. Tools for Single Cell Genomics. [cited 7 Oct 2022]. Available: https://satijalab.org/seurat/

102. Xi Peng Institute for Infocomm Research, Singapore, Shijie Xiao Nanyang Technological University, Singapore and OmniVision Technologies Singapore Pte. Ltd, Jiashi Feng National University of Singapore, Singapore, Wei-Yun Yau Institute for Infocomm Research, Singapore, Zhang Yi College of Computer Science, Sichuan University, Chengdu, P. R. China. Deep subspace clustering with sparsity prior. In: Guide Proceedings [Internet]. [cited 7 Oct 2022]. Available: https://dl.acm.org/doi/10.5555/3060832.3060890

103. Mahdizadehaghdam S, Panahi A, Krim H, Dai L. Deep Dictionary Learning: A PARametric NETwork Approach. IEEE Trans Image Process. 2019. doi:10.1109/TIP.2019.2914376

104. IEEE Xplore Full-Text PDF: [cited 7 Oct 2022]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7779008

105. Vanika Singhal Indraprastha Institute of Information Technology, Delhi, India, Angshul Majumdar Indraprastha Institute of Information Technology, Delhi, India. Majorization Minimization Technique for Optimally Solving Deep Dictionary Learning. Neural Process Letters. 2018 [cited 7 Oct 2022]. doi:10.1007/s11063-017-9603-9

106. [No title]. [cited 7 Oct 2022]. Available: http://proceedings.mlr.press/v75/yarotsky18a/yarotsky18a.pdf

107. On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization. [cited 7 Oct 2022]. Available: https://ieeexplore.ieee.org/document/4359171

108. Blakeley P, Fogarty NME, Del Valle I, Wamaitha SE, Hu TX, Elder K, et al. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. Development. 2015;142: 3613.

109. Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, et al. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. Cell Stem Cell. 2015;17: 471–485.

110. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol. 2013;20: 1131–1139.

111. Tools for Single Cell Genomics. [cited 7 Oct 2022]. Available: https://satijalab.org/seurat/

112. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45: D833–D839.

113. McKusick VA. Mendelian Inheritance in Man and its online version, OMIM. Am J Hum

Genet. 2007;80: 588–604.

114. Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinformatics. 2015;16: 1–17.

115. Quan C, Ren F. Gene–disease association extraction by text mining and network analysis. Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi). 2014. pp. 54–63.

116. Zhou J, Fu B-Q. The research on gene-disease association based on text-mining of PubMed. BMC Bioinformatics. 2018;19: 37.

117. Wan F, Zeng J (michael). Deep learning with feature embedding for compound-protein interaction prediction. bioRxiv. 2016. p. 086033. doi:10.1101/086033

118. A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining. [cited 7 Oct 2022]. Available: https://ieeexplore.ieee.org/document/8730332

119. Kim J, So S, Lee H-J, Park JC, Kim J-J, Lee H. DigSee: Disease gene search engine with evidence sentences (version cancer). Nucleic Acids Res. 2013;41: W510–7.

120. Zhou K, Zhang S, Wang Y, Cohen KB, Kim JD, Luo Q, et al. High-quality gene/disease embedding in a multi-relational heterogeneous graph after a joint matrix/tensor decomposition. J Biomed Inform. 2022;126. doi:10.1016/j.jbi.2021.103973

121. Yang K, Wang R, Liu G, Shu Z, Wang N, Zhang R, et al. HerGePred: Heterogeneous Network Embedding Representation for Disease Gene Prediction. IEEE J Biomed Health Inform. 2019;23: 1805–1815.

122. Yang C, Xiao Y, Zhang Y, Sun Y, Han J. Heterogeneous Network Representation Learning: A Unified Framework with Survey and Benchmark. 2020 [cited 7 Oct 2022]. doi:10.48550/arXiv.2004.00216

123. Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature. 2019;571: 95–98.

124. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47: D1005–D1012.

125. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer. 2004;91: 355–358.

126. [No title]. [cited 7 Oct 2022]. Available: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1286&context=asc_papers

127. BioSentVec: creating sentence embeddings for biomedical texts. [cited 7 Oct 2022]. Available: https://ieeexplore.ieee.org/document/8904728

128. scikit-learn. [cited 7 Oct 2022]. Available: https://scikit-learn.org/stable/

129. LIBSVM -- A Library for Support Vector Machines. [cited 7 Oct 2022]. Available: https://www.csie.ntu.edu.tw/~cjlin/libsvm/

130. Sharma N. XGBoost. The Extreme Gradient Boosting for Mining Applications. GRIN Verlag; 2018.

131. Hosmer DW Jr, Lemeshow S. Applied Logistic Regression. John Wiley & Sons; 2004.

132. Random decision forests. [cited 7 Oct 2022]. Available: https://ieeexplore.ieee.org/document/598994

133. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ. 1998;32: 2627–2636.

134. Perkins J. Python Text Processing with Nltk 2.0 Cookbook: Lite. Packt Publishing Ltd; 2011.

135. Wei C-H, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. 2019;47: W587–W593.

136. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45: 580–585.

137. Maag JLV. gganatogram: An R package for modular visualisation of anatograms and tissues based on ggplot2. F1000Res. 2018;7. doi:10.12688/f1000research.16409.2

138. De Fusco M, Marconi R, Silvestri L, Atorino L, Rampoldi L, Morgante L, et al. Haploinsufficiency of ATP1A2 encoding the Na+/K+ pump alpha2 subunit associated with familial hemiplegic migraine type 2. Nat Genet. 2003;33: 192–196.

139. Hartl D, Griese M, Nicolai T, Zissel G, Prell C, Reinhardt D, et al. A role for MCP-1/CCR2 in interstitial lung disease in children. Respir Res. 2005;6: 93.

140. Trochet D, O'Brien LM, Gozal D, Trang H, Nordenskjöld A, Laudier B, et al. PHOX2B genotype allows for prediction of tumor risk in congenital central hypoventilation syndrome. Am J Hum Genet. 2005;76: 421–426.

141. DISEASES. [cited 7 Oct 2022]. Available: https://diseases.jensenlab.org/About

142. Mahley RW. Apolipoprotein E: from cardiovascular disease to neurodegenerative disorders. J Mol Med . 2016;94. doi:10.1007/s00109-016-1427-y

143. Karch CM, Cruchaga C, Goate AM. Alzheimer's disease genetics: from the bench to the clinic. Neuron. 2014;83. doi:10.1016/j.neuron.2014.05.041

144. Singh N, Cohen PL. The T cell in Sjogren's syndrome: force majeure, not spectateur. J Autoimmun. 2012;39. doi:10.1016/j.jaut.2012.05.019

145. Chan VS-F, Tsang HH-L, Tam RC-Y, Lu L, Lau C-S. B-cell-targeted therapies in systemic lupus erythematosus. Cellular and Molecular Immunology. 2013;10: 133.

146. Yap H-Y, Tee SZ-Y, Wong MM-T, Chow S-K, Peh S-C, Teow S-Y. Pathogenic Role of Immune Cells in Rheumatoid Arthritis: Implications in Clinical Treatment and Biomarker Development. Cells. 2018;7. doi:10.3390/cells7100161

147. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44. doi:10.1093/nar/gkw377

148. Davies HR, Hodgson K, Schwalbe E, Coxhead J, Sinclair N, Zou X, et al. Epigenetic modifiers DNMT3A and BCOR are recurrently mutated in CYLD cutaneous syndrome. Nat Commun. 2019;10: 1–9.

149. Li J, Zheng Y, Li X, Dong X, Chen W, Guan Z, et al. UCHL3 promotes proliferation of colorectal cancer cells by regulating SOX12 via AKT/mTOR signaling pathway. Am J Transl Res. 2020;12. Available: https://pubmed.ncbi.nlm.nih.gov/33194042/

150. Kawahara H, Minami R, Yokota N. BAG6/BAT3: emerging roles in quality control for nascent polypeptides. J Biochem. 2012;153: 147–160.

151. Kim J, Kim JJ, Lee H. An analysis of disease-gene relationship from Medline abstracts by DigSee. Sci Rep. 2017;7. doi:10.1038/srep40154

152. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of proteins and transcripts in single cells. Nat Biotechnol. 2017;35: 936–939.

153. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. 2018;19: 1–12.