A single cell RNA-Seq based aging clock for human neurons

Qiao Su
Department of Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213

Mirudhula Mukundan
Department of Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213

1 Introduction

Aging is a biological process which affects every single cell in the body, but the exact molecular mechanisms remain unknown. In order to decipher this mechanism, we can study which gene features are predictive of chronological age. Phenotypes such as aging are usually studied at the level of bulk data, which comes from aggregated expression data from multiple cells.

Past studies have shown some success predicting chronological age from bulk human and mouse gene expression data using elastic net models (1). However, these bulk tissue measurements do not explain how heterogeneous populations of cells differ in terms of their gene expression as aging progresses. Fortunately, single-cell sequencing technology allows us to examine disease biology in unprecedented detail.

However, there are also new computational challenges involved in processing and interpreting this high dimensional expression data. In addition to biological noise from the stochastic process of mRNA production, there is technical noise due to the low amounts of genetic material that can be captured from a single cell. Therefore, deep learning classifiers may be especially well suited to this challenge.

We hypothesize that chronological age can be predicted from features of single-cell expression data using deep learning methods. Towards this goal, we aim to compare deep learning models for predicting age from a real biological dataset, which in this paper is taken from brain samples of human control patients of a study of schizophrenia (10). Furthermore, we try out different dimensional reduction methods to compare the result on performance.

1.1 Introduction to NMF data decomposition

The interpretation of scRNA-seq data requires methodological innovations. Matrix factorization approaches can find a simplified and thus more interpretable representation of an expression matrix which can also be less costly for a deep learning model to train. For example, exploratory analysis is usually assisted by unsupervised matrix factorization approaches such as principal components analysis (PCA) or t-distributed stochastic neighbour embedding (t-SNE). Non negative matrix factorization (NMF) is another popular matrix factorization approach with the advantage that it has been shown to be able to discover biologically meaningful gene expression programs as latent factors compared to PCA and t-SNE (8). This approach has been successfully used for cell type markers in the past, but it has not yet been applied towards deciphering gene expression programs such as human aging. We will attempt to infer the top 100 latent variables using NMF and check their correlation with Age.

1.2 Introduction to Variational Autoencoder

Studying patterns in the level of transcripts in the scRNA-seq data can lead to identifying key molecular signatures responsible for the aging process. But, this data is highly complex in nature and

10617 Class Project (2022).

poses great challenges in both regression analysis and feature selection. Following this, a very high features-to-examples ratio, which is a widely prevalent characteristic of biological and medical data, makes the model close to uninterpretable as well. Scientists have come up with ways to compress large datasets in order to identify and consolidate across multiple latent dimensions to capture key gene expression representations in a quick and efficient manner (14). Previous studies have worked with novel techniques that involve variational autoencoders (VAE) to compress RNA-seq data for extracting latent dimensions (13) or for performing dimensional reduction in scRNA-seq data related to cancer (5). But, no study has yet been conducted on the application of VAEs on ageing related transcriptomic data. Therefore, this project will explore how regularization between a series of linear layers of an autoencoder will help in identifying good latent dimensions for efficient regression.

1.3 Introduction to Denoising Autoencoder

One defining characteristic of gene expression data is a low ratio of signal to noise. Therefore, we sought to find a model which would perform well with noisy data. In addition to the standard VAE implementation, previously a new approach called DAE (denoising autoencoder) has been explored ((6)), however it has only been applied towards image data. The authors found that adding noise to the VAE when training on the MNIST dataset improves performance. We will explore whether this same approach may help us improve performance on our gene expression data as well.

2 Related Work

A comparative study between chronological age and biological age is useful in understanding aging as a phenomenon explained by a set of biomarkers. Recently, studies have shown to indicate that the internal aging clock can be explained by using gene expression data (2; 3). Some studies have utilized this to build models for prediction of age using Genotype-Tissue Expression (GTEx) profile from multiple human tissues using an elastic net algorithm (16; 12) to account for sparsity in data. Another study conducted age prediction analysis in zebrafish with the help of a simple Multilayer Perceptron Model (MLP) with one linear layer, two ReLU activation functions, and a softmax layer to predict the biological age of the zebrafish into three discrete bins indicating stage of age in life (11). In our project, we incorporated single-cell gene expression data from human neurons to predict biological age of the subject using three different techniques: (a) Deep Non-negative Matrix Factorization with, (b) Variational Autoencoder, and (c) De-noising Autoencoders.

3 Background

3.1 Data

For this project, we are working with gene expression data at the single-cell level, extracted from human neurons. The entire dataset consists of 189432 human neurons from 69 controls, containing a distribution of ages from 25 to 94 years (Fig S1). To facilitate the fit of the deep learning model, we applied feature scaling of both the target variable as well as z-scoring of the feature matrix using sklearn StandardScaler.

3.2 Splitting the Data

The data was split into three parts: (a) Train set with 50 subjects, (b) Validation set with 9 subjects, and (c) Test set with 10 subjects. This was done in order to further reduce bias during hyperparameter tuning. Therefore, all initial and tuning experiments were conducted on train and validation set. Finally, all results are reported based on the never-before-seen test set.

3.3 Baseline Model

Initially we implemented an elastic net model baseline model as well as a baseline Multi-layer Perceptron (MLP) with varied architecture. We then obtained the predicted ages and averaged the predictions from all cells of a each subject. This was taken as the final predicted age of the subject. We calculated the Mean Absolute Error (MAE) was calculated for the pair of predicted and actual age. The models used are as listed below:

1. An elastic net model with parameters tuned by 5-fold cross validation, for the purposes of comparison with previously published elastic net models.
2. An MLP model was implemented with sequence of linear and batch normalization layers, ReLU as the activation function, ADAM as the optimizer, dropout probability of 0.4, and a batch size of 4096. We conducted a small hyperparameter tuning protocol by varying the numbers of layers, number of neurons in each layer and the learning rate, in order to find the best architecture with minimum MAE between the predicted (validation) ages and the actual ages. The tuning results are as given in Table 1. We observed that the MAE was least when we had 3 layers with 5000 neurons in each layer, and with a learning rate of 0.001. Therefore, this was taken as the final baseline model architecture.
3. Following the second model, we deduced that a dimensional reduction may further improve the predictions, especially given the large number of features in our data. So, we built a model based on Non-negative Matrix Factorization (NMF)

The Linear Regression plots of the final predicted and actual age in the test set for all baseline models is shown in Fig. 1. We observed that our MLP model and NMF dimensionality reduction in fact performed worse than the linear model, with no improvement beyond 30 epochs.

Model	Hidden Layer Dim	#Layers	Learning Rate	MAE
1	[500,500]	2	0.001	8.44
2	[1000,1000]	2	0.001	9.11
3	[5000,5000,5000]	3	0.001	7.78
4	[5000,5000,5000]	3	0.0001	8.67

Table 1: Results of hyperparameter tuning for the MLP model.

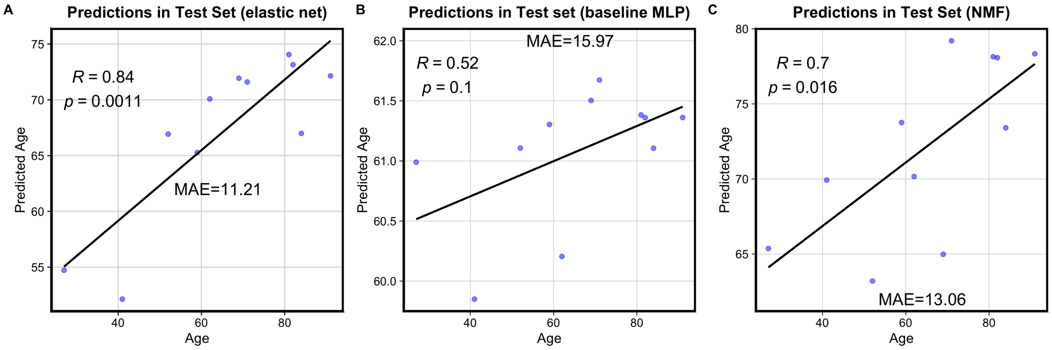

Figure 1: Baseline Results and NMF
. From L to R: (a) plot of elastic net model results, (b) plot of baseline MLP model results, (c) plot of baseline MLP results with NMF dimensional reduction.

4 Methods

Our goal was to reduce the number of dimensions in order to derive meaningful latent representations of the data, which would help us better predict the subject ages. For this, we implemented three different techniques, as listed below. In all our models, we implement the same basic MLP architecture for predicting age that we got from the best baseline MLP model, i.e. each of the technique listed below will have a sequence of Linear, Batch Normalizarion, ReLU and Dropout layers, with ADAM as the optimizer at 0.001 learning rate and 4096 batch size. This is done in order to compare between multiple techniques.

4.1 SVD

For the purpose of exploratory data analysis, we performed SVD on the entire dataset with the python sklearn package and default parameters, using a dimension of 100. We first checked whether the data contained strong variance across Age by performing singular value decomposition (SVD) on the dataset, which is a quick and computationally tractable alternative to PCA. The top pearson correlation of SVD components with Age was found to be 0.43 (Supplementary Fig S2). Therefore the components did not exhibit a particularly high correlation with the target variable.

4.2 Non-negative Matrix Factorization

We performed NMF on the raw training and validation split of the data with the python sklearn package and default parameters, using a dimension of 100. The NMF model was fit to the training and validation split and applied to scale the test data.

4.3 Variational Autoencoder (VAE)

Previous studies have worked with novel techniques that involve variational autoencoders (VAE) to compress RNA-seq data for extracting latent dimensions and for dimensionality reduction (13; 5). In this section, we implemented two different variations of VAE. This model has the advantage of having a variational inference which helps achieve a more scalable precise latent representation. Another advantage is the ease with which we can model the data by comparing appropriate likelihood functions.

4.3.1 Standard Variational Autoencoder

In this method, we applied the generative model to the scaled features and target variable by z-scoring them respectively with respect to the training data only. VAEs work by introducing stochastic latent variables on which the generative process is conditioned. This usually forms the likelihood function of the model, with the probability of the latent variables being the prior. For all datapoints, we sum over the log-likelihoods and maximum likelihood in order to estimate the parameters. In the standard VAE, we use a standard multivariate Gaussian distribution as both the likelihood and the prior, described as:

$$\begin{gathered}\mathrm{p}_\theta(x|z) = \mathcal{N}(\mu, \sigma^2),\\ \mathrm{p}_\theta(z) = \mathcal{N}(0, I),\end{gathered}$$

where the data x is represented by a continuous distribution, and z is the latent variable. We can use the reparameterization trick in order to estimate our latent variable by using single linear layer, which will get trained for weights and biases along with the rest of the network. This is the model derived from the original version of VAE by Kingma *et al.*(7) that is widely prevalent in the field of Deep Learning.

4.3.2 Poisson Variational Autoencoder

Athough the previous model is still a strong model for different kinds of data, we realize that our original unstandardized data is comprised mainly of discrete counts of the transcripts produced that make up the gene expression. So, by z-scoring or normalizing them, we lose out on the data's characteristic *sequencing depth*. In addition to this, our data is extremely sparse with a large number of zeros. This is a vital piece of information when we model the data. Therefore, in order to model these discrete count values, we deduced that a Poisson distribution as the likelihood function would capture the true image of the data. Therefore, our likelihood would now change to the following:

$$\begin{gathered}\mathrm{p}_\theta(x|z) = f(\lambda_\theta(z)),\\ \mathrm{p}_\theta(z) = \mathcal{N}(0, I),\end{gathered}$$

This method closely follows the implementations of Christopher *et al.* and Zhao *et al.*, for their implementations of VAE on sparse data (5; 15). Note that the prior distribution and reparameterization trick was done in a way similar to the standard VAE, and we forgoed the standardization step at the beginning of the experimental run.

4.3.3 Autoencoder Architecture

Both the autoencoders implemented above had similar architectures. An example of the architecture with the standard VAE is given in Fig. 2. The same is used in the case of the Poisson VAE, with the

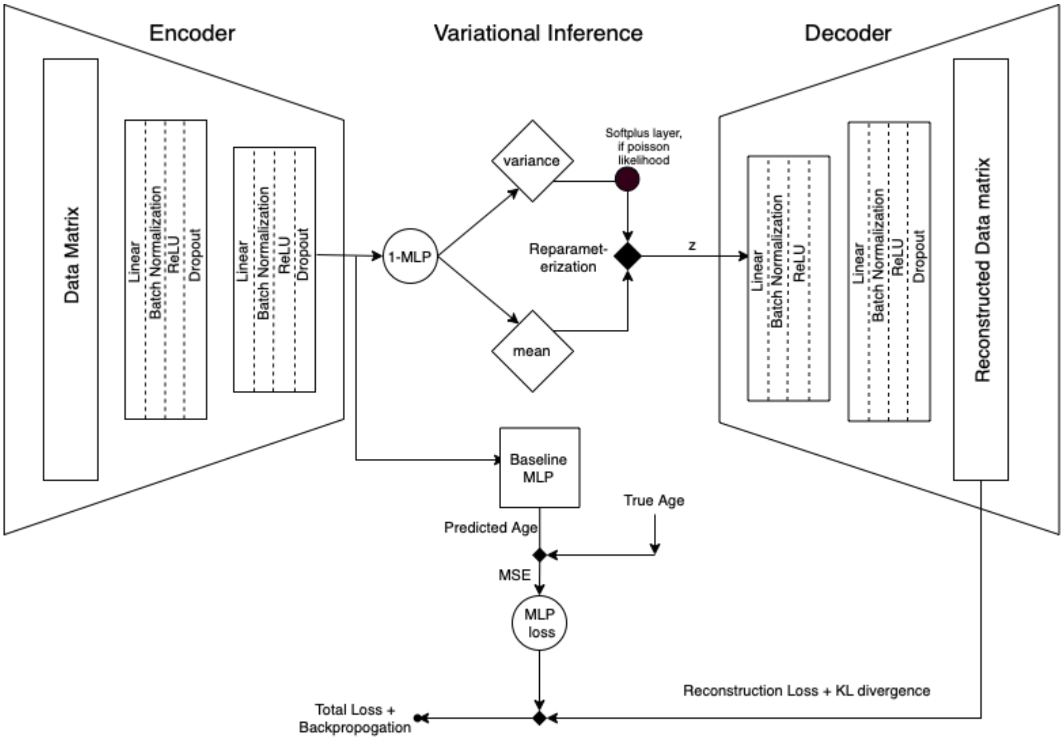

Figure 2: Autoencoder Architecture used in VAE.

exception of the changed likelihood function and the addition of a softplus layer after calculating the variance, in order to get positive values. In order to avoid posterior collapse, where the autoencoder does not learn anything informative, we implemented the architecture with Kullback–Leibler divergence (KL) loss annealing (4). This collapse could occur when loss from the KL minimizes too fast, because of which the reconstruction losses are still very high and the autoencoder does not learn the data efficiently beyond the collapse. Therefore, we set a scheduler for the introduction of the KL loss from the 20th epoch, after which there is an exponential increase in the percentage of contribution of the KL loss for the next 10 epochs, and then full contribution for another 10 epochs. After this, the loss is reset and the same cycle starts again. We also observed that the data seemed to have many plateaus and steeps. In order to cover all geometry, we also implemented a learning rate scheduler involving cosine annealing with warm restarts (9) where the learning rate decays exponentially for 20 epochs before restarting again.

4.3.4 Denoising Criterion with Variational Autoencoder

Since VAE seemed to be producing good results, we thought we could add a denoising criterion to the Standard VAE model, to observe if this would help mitigate the noise in the data, which is a characteristic of biological data. Borrowing from the data handling of the base VAE model, we scaled features with a classifier fit to the train data matrix only but applied to the test and validation datasets. We used the same neural network architectures which were chosen to be optimal in parameter tuning of the base VAE. Additionally, the DAE differs from the base VAE because a gaussian noise is added to the inputs of the autoencoder. In the previously published implementation from (6), empirical studies of the DAE showed that introducing a small level of noise surpassed the performance of vanilla VAE but introducing too much noise will lead to worse performance than the vanilla VAE. Therefore we introduced two levels of Gaussian noise as indicated in table 4, from 0.5-5.

5 Results & Discussion

In this section, we report the final results of each of our model, followed by a short discussion about the implementation and future work for the respective models.

5.1 NMF

After the hyperparameter tuning on the validation set, we checked whether the baseline neural network architecture trained with the 100 NMF components could provide an improvement in performance on the test set (Figure 1c compared to Figure 1b). In this figure, we saw that the dimensional reduction seemed to increase Pearson correlation from 0.52 to 0.7 while decreasing MAE to 13.08 from 15.97, in general improving the signal explained. This makes sense as we believed based on the literature that the NMF may provide an improvement in representing this biological data compared to methods such as PCA and SVD. However, the MAE is still higher than the performance of the elastic net (Figure 1a). This suggests that NMF dimensional reduction provides a marginal improvement over the baseline model. This may make sense given that NMF is a dimensional reduction similar to SVD and we did not find that SVD does not produce components significantly correlated with the variable Age. It is possible that the complexity of the data transformation is not enough to capture the signal of the target variable, which is relatively subtle.

5.2 Variational Autoencoder

5.2.1 Results

A number of experiments were conducted in order to get the best set of hyperparameters. An initial set of experiments was run on the Standard VAE without dropouts. Note that all tuning was done on the validation set of the data, and not the test set. The results of tuning are as listed in Table 2. All experiments were run with ADAM optimizer and learning rate 0.001.

The results of the Standard VAE did not seem quite satisfactory as we were not able to beat the baseline model in any of the models. Even the 4th model, which appears to be close to the baseline might be arbitrary since the very next model with a slight latent dimension increase by 10 does not seem to do as well as this model.

The next set of experiments included using the Poisson VAE. Initially, we conducted this model without dropouts, but including them seemed be giving a nice gradual decrease in the loss. The results of the tuning for this model are as listed in Table 3. In this modified VAE, we were in fact able to beat the baseline on all our runs of the experiment, with the MAE being below 10. This model seemed to be learning better and better as we increased the number of epochs, but after 1000 epochs, the learning seemed to have become even slower, so we stopped our learning at 1000 epochs. From this tuning, we were able to decide our final model hyperparameters to be model number 3 from the table. This model was used to predict the ages in the test set, and we were able to achieve an MAE of 7.55, which was lesser than the validation result. This can confirm the model does perform well and can beat the baseline on unseen data as well. The Linear Regression plot for the pair of predicted and actual ages is shown in Fig.3.

5.2.2 Discussion

Summarizing the results of VAE, the performance of the model seems to vary with how we decide to model the data. Especially for a data like single-cell RNA expression, where the matrices are very sparse and complexity is high, it is indeed tough to model the data. In this section, we were able to successfully model the data better than the baseline results, not only because of the final MAE value, but also because the results of the model seem more stable than during hyperparameter tuning of the baseline, where it does not seem easy to predict how the trends in the tuning will affect the predictions. Our poisson model, relatively, was able to put forth a better outcome, and it may be possible to increase the precision if we run it for more number of epochs, albeit with a careful eye on overfitting.

It is also possible to use other likelihood functions like Negative Binomial, which is known to model over-dispersed data very well. Some VAE models for scRNA also make use of a zero-inflated Poisson and Negative Binomial likelihoods, which would help in keeping the effect of zeros on the probabilities under control (5). Another approach with negative binomial likelihood that could help capture the non-linearities in the data as well as understand the relationship between the latent space and sample space is to construct the prior by making use of the original features (15). This is important since for modeling any biological data using models that convert the original data into latent dimensions, it is important to get back the original features and their relevance in the classification or regression. This, for example, can help us identify key genes or molecular signatures that help us understand the biological problem at hand.

Model	Epochs	MLP Dimensions	Latent Dimensions	Autoencoder Dimensions	Dropout	MAE
1	100	[20,20]	20	[300,100]	No	15.78
2	100	[20,20]	30	[300,100]	No	13.44
3	100	[20,20]	40	[300,100]	No	14.67
4	100	[20,20]	50	[300,100]	No	12.67
5	100	[20,20]	60	[300,100]	No	14.22
6	100	[20,20]	70	[300,100]	No	14.33
7	100	[20,20]	50	[600,150,75]	No	13.0
8	100	[30,20]	150	[512,256]	No	13.0
9	200	[20,20]	60	[300,100]	Yes	14.22

Table 2: Results of hyperparameter tuning for Standard VAE.

Model	Epochs	MLP Dimensions	Latent Dimensions	Autoencoder Dimensions	Dropout	MAE
1	200	[20,20]	100	[300,150]	Yes	9.56
2	500	[20,20]	100	[300,150]	Yes	8.67
3	**1000**	**[20,20]**	**100**	**[300,150]**	**Yes**	**8.33**

Table 3: Results of hyperparameter tuning for Poisson VAE.

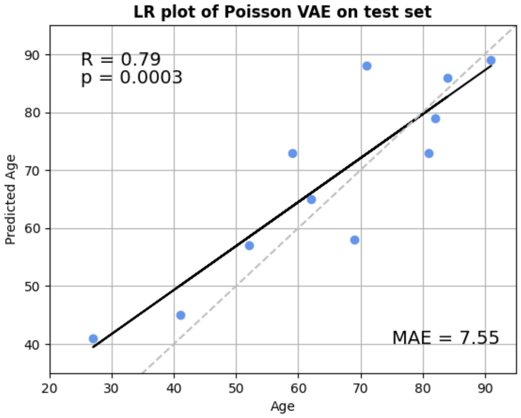

Figure 3: LR plot for Poisson VAE with test set.

5.3 Denoising Criterion with Variational Autoencoder

5.3.1 Results

Hyperparameter tuning was conducted on the validation set, and is listed in Table 4. Since model #8 from the Standard VAE tuning seemed be giving good results from amongst the others, we started our tuning for this model with the same architecture as this. We added gaussian noise scaled to be either small (0.5) or large (5) relative to the maximum count. Based on the results of the tuning, adding noise does not improve the performance of the DAE relative to the Standard VAE, however DAE Model #1 with the maximum amount of noise seems to perform marginally better than the other hyperparameter configurations. Finally, we fit DAE Model #1 to the training and validation data and predicted on test data. The predicted ages from the test set was compared against the actual ages in Fig.4. An MAE of 16.09 on the test set suggests that the DAE underperforms relative to the VAE.

Model	Epochs	MLP Dimensions	Latent Dimensions	Autoencoder Dimensions	Noise	Dropout	MAE (validation)
1	**100**	**[20,20]**	**100**	**[512,256]**	**5**	**No**	**14.07**
2	100	[20,20]	100	[512,256]	0.5	No	14.43
3	100	[30,20]	100	[512,256]	0.5	No	14.14

Table 4: Results of hyperparameter tuning for Denoising VAE.

5.3.2 Discussion

We believe that there are two main reasons why the DAE did not work out with these hyperparameter configurations. Firstly, the Poisson VAE worked with great success and we believe that the next step to extend our DAE would be to incorporate the poisson model in the place of the gaussian model for these sparse counts. Secondly, we had not done a complete hyperparameter search for the DAE due to time and compute constraints especially to find the optimal noise level as the model is very sensitive to this. In addition to this, the reference publication (6) has multiple suggestions that could be incorporated into future work, for example increasing sample size and introducing a more realistic model than the Gaussian for the noise in the counts.

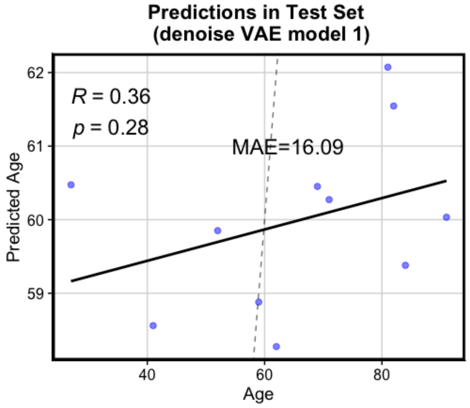

Figure 4: LR plot for Denoising criterion with Standard VAE on test set.

6 Conclusion

In this project, we applied a deep learning model to the task of predicting Age of patients based upon their single cell count data. However, the deep learning MLP was found to underperform an elastic net model. We then checked whether three different dimensional reduction techniques could improve the performance of a baseline MLP model with fixed architecture. Out of these, we found that NMF and VAE both had improved performance relative to the baseline, but the DAE did not. However using NMF components did not improve the performance of the model relative to a linear elastic net model, potentially due to the low complexity of this dimensional reduction. Therefore out of the three dimensional methods, the poisson VAE model seemed to have performed best with MAE below that of the baseline linear model.

7 Supplemental Figures

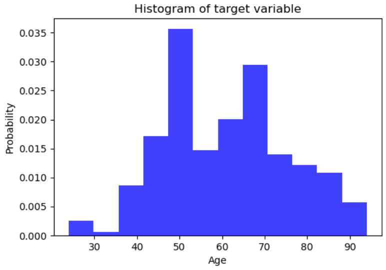

Figure 1: Distribution of target variable (Age) across 189432 individual cells measured.

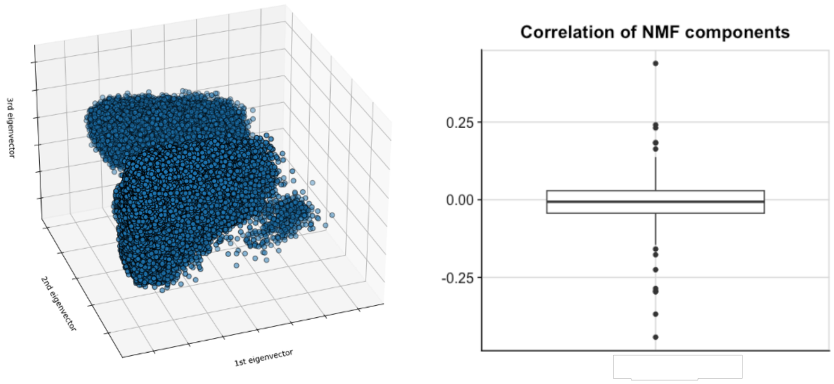

Figure 2: First 3 components from SVD (left) and boxplot of the correlation of all components to Age (right).

References

[1] Bergsma, T., and Rogaeva, E. DNA Methylation Clocks and Their Predictive Capacity for Aging Phenotypes and Healthspan. *Neuroscience Insights 15* (2020).

[2] De Magalhães, J. P., Curado, J., and Church, G. M. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics 25*, 7 (2009), 875–881.

[3] Frenk, S., and Houseley, J. Gene expression hallmarks of cellular ageing. *Biogerontology 19*, 6 (2018), 547–566.

[4] Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., and Carin, L. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145* (2019).

[5] Grønbech, C. H., Vording, M. F., Timshel, P. N., Sønderby, C. K., Pers, T. H., and Winther, O. scvae: variational auto-encoders for single-cell gene expression data. *Bioinformatics 36*, 16 (2020), 4415–4422.

[6] Im, D. J., Ahn, S., Memisevic, R., and Bengio, Y. Denoising criterion for variational auto-encoding framework. *31st AAAI Conference on Artificial Intelligence, AAAI 2017* (2017), 2059–2065.

[7] Kingma, D. P., and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[8] Kotliar, D., Veres, A., Nagy, M. A., Tabrizi, S., Hodis, E., Melton, D. A., and Sabeti, P. C. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife 8* (jul 2019).

[9] Loshchilov, I., and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

[10] Ruzicka, B., Mohammadi, S., Davila-Velderrain, J., Subburaju, S., Tso, R., Hourihan, M., and Kellis, M. Single-Cell Dissection of Schizophrenia Reveals Neurodevelopmental-Synaptic Link and Transcriptional Resilience Associated Cellular State. *Biological Psychiatry 89*, 9 (2021), S106.

[11] Singh, S. P., Janjuha, S., Chaudhuri, S., Reinhardt, S., Kränkel, A., Dietz, S., Eugster, A., Bilgin, H., Korkmaz, S., Zararsiz, G., et al. Machine learning based classification of cells into chronological stages using single-cell transcriptomics. *Scientific reports 8*, 1 (2018), 1–12.

[12] Wang, F., Yang, J., Lin, H., Li, Q., Ye, Z., Lu, Q., Chen, L., Tu, Z., and Tian, G. Improved human age prediction by using gene expression profiles from multiple tissues. *Frontiers in Genetics 11* (2020), 1025.

[13] Way, G. P., and Greene, C. S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium* (2018), World Scientific, pp. 80–91.

[14] Way, G. P., Zietz, M., Rubinetti, V., Himmelstein, D. S., and Greene, C. S. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. *Genome biology 21*, 1 (2020), 1–27.

[15] Zhao, H., Rai, P., Du, L., Buntine, W., Phung, D., and Zhou, M. Variational autoencoders for sparse and overdispersed discrete data. In *International Conference on Artificial Intelligence and Statistics* (2020), PMLR, pp. 1684–1694.

[16] Zou, H., Hastie, T., et al. Addendum: regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B 67*, 5 (2005), 768–768.