OXFORD

Gene compression using Machine Learning for Cancer type Classification

Rushi Bhatt [1], Ronak Kaoshik [1] and Shruti Mohanty [1]

[1] University of California, Los Angeles, 90024, USA

Abstract

Motivation: Compression algorithms are applied to gene data to capture latent features that represent different biological sources of variation. These representations reveal valuable information about the genomic space to generate hypotheses that would be difficult to observe in the original feature space. These representations can be used to develop robust models for performing downstream tasks like cancer type classification, predict gene alteration status, etc. Choosing a single latent dimensionality and compression algorithm is a significant step, and this relies on certain heuristics as well as the targeted task. In this project we focus on the task of cancer type classification using RNAseq data. We improvise significantly on the existing works by using ensemble methodology for compression and neural network based learning techniques for classification.

Results: We trained four compression algorithms - PCA, ICA, NMF, and VAE across a range of latent dimensionalities on The Cancer Genome Atlas PanCanAtlas(TCGA) dataset. They were evaluated across the metrics - Pearson correlation, reconstruction cost, and accuracies for logistic regression and neural network classifiers. From this, we concluded various optimal dimensionalities for each of the algorithms and later leveraged it in the ensemble method. When this is trained using Vanilla Neural Network model we learn the best biological representations achieving 97.5 % accuracy for the prediction of cancer types.

Availability: Code to perform all analysis and generate the results provided in this report is present in https://github.com/RushiBhatt007/Gene-compression-and-Cancer-type-classification .

Contact: rushibhatt@g.ucla.edu ,ronak42@g.ucla.edu , shrutimohanty@g.ucla.edu

1 Introduction

In recent times, high-throughput sequencing technologies have made genome sequencing relatively cheaper and has led to processing of genomes from a wide range of subjects on a very large scale. With the collection of massive amounts of data, there is a need to develop sophisticated techniques that are able to compress and extract the rich information from these sequences in an effective manner. Gene sequence compression algorithms are being developed for this specific purpose and have started to produce promising results.

There are simple algorithms like Principal Component Analysis, Independent Component Analysis and Non-negative Matrix Factorization which make use of linear expressions to reveal the relations and expressions of the compressed feature set. Then there are non linear methods which make use of various advanced concepts such as Deep Learning Networks, Reconstruction networks, to obtain an even better representation of the feature set.

In recent research works these algorithms have been thoroughly explored in terms of the latent space dimensions, however, the effectiveness of compression in downstream tasks like cancer type classification and others is not studied thoroughly. We present a study that focuses on multi-class cancer type classification using data obtained from various gene compression techniques and we also propose our own approach which outperforms all the existing methods.

2 Approach

We implemented the pipeline shown in Figure 1 inspired from Way, G.P. *et al*. on TCGA dataset using four different algorithms – PCA, ICA, NMF, and VAE. From these compressed features into latent dimensionalities, we've calculated performance metrics like reconstruction loss, Pearson correlation and Spearman correlation. Since there is no generalized latent space that captures all the biological features from RNASequence, our focus in this work is mainly on the task of Cancer-type Classification. For the classification task, we have initially implemented a logistic regression model to get a baseline performance for this task and found out that these features in isolation do not perform that well. Therefore, to further improve our performance, we have implemented a neural network-based classifier using an ensemble of the elements from all four algorithms, which achieved an accuracy of 97.5% for 12-class classification.

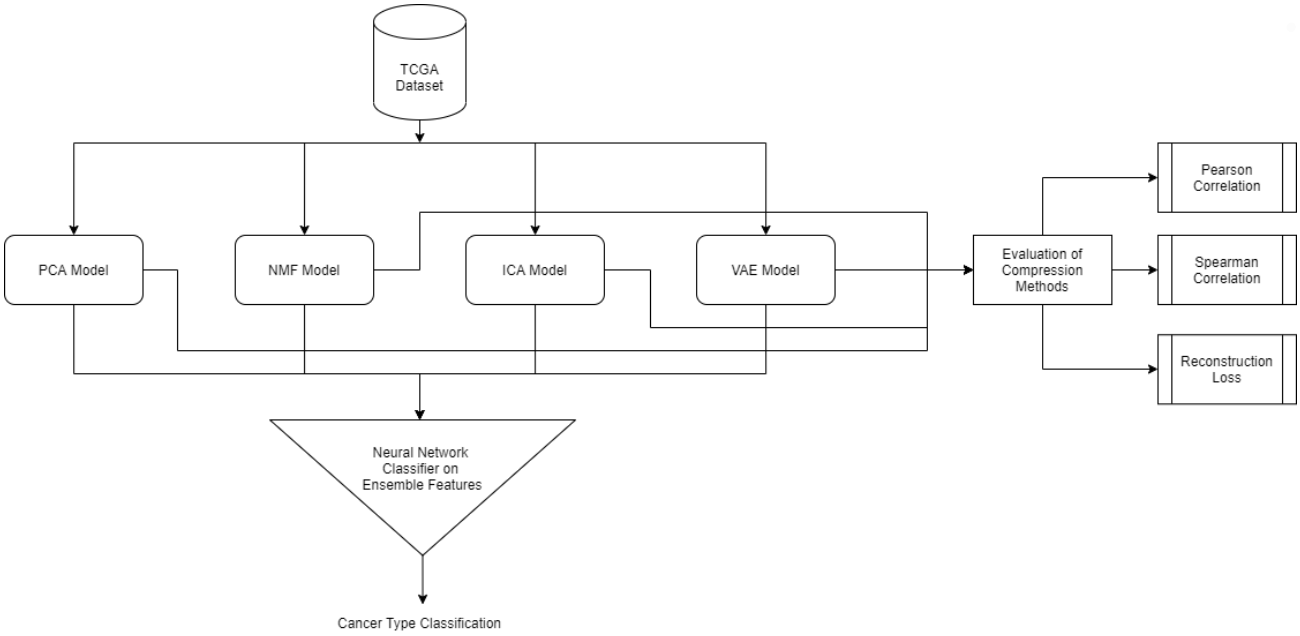

Fig. 1. Proposed model pipeline

3 Methods

This section explains the model pipeline in detail.

3.1 Gene Compression

We applied one-step linear compression algorithms like Principal Component Analysis (PCA), Independent Component Analysis(ICA), and Non Negative Matrix Factorisation (NMF). These are legacy algorithms that have been widely accepted in dimensionality reduction.

PCA will identify a unique and deterministic solution that represents compressed features with a decreasing amount of variance. The directions identified by PCA are orthogonal to one another and linearly uncorrelated. As explained in Jolliffe *et al*., 2016, often in PCA selecting the top most features will represent the entire dataset, however, there can be circumstances where the last few are of interest for outlier detection. So, selecting features across multiple latent dimensions is a more generalized representation model.

Unlike PCA, ICA and the other algorithms have no inherent ordering present for their feature sets. From Kairov *et al*., 2017, ICA will define a new coordinate system in the multidimensional space such that the distributions of the data point in new axes are as mutually independent as possible. The mixed-signal separation works extremely well in ICA when the subcomponents are derived from a non-gaussian distribution. ICA has been widely applied for the analysis of transcriptomic data for blind separation of biological, and technical factors affecting gene expression data. As explained in Frigyesi *et al*., 2008, Non-negative matrix factorization (NMF) is a relatively new approach to analyze gene expression that models data by additive combinations of non-negative basis vectors. Gene expression data is split into a product of two non negative matrices W and H. The k columns of W are the basis vectors using which factorization is performed in the genetic space.

Linear methods have been applied to the large transcriptomic compendium in gene expression data to reveal the influence of copy number alterations, identify coordinated transcriptional programs, and estimate cell-type proportion in bulk tissue samples. Nonlinear methods such as the VAE reveal latent signals characterizing oxygen exposure, cancer subtypes, and drug response. So using an ensemble of these algorithms will give a good reduced set feature representation model for biological signals.

VAE is a neural network-based compression and data generation technique to learn meaningful latent dimensions. It is based on an autoencoder framework, consisting of an encoder where to input is projected into a latent dimensional space and a decoder, which reconstructs the input from this lower dimension space. Traditional autoencoders learn by minimizing the reconstruction error. However, VAE learns the distribution of these latent features by their mean and standard deviation. Moreover, the reconstruction loss of VAE is a combination of Kullback-Leibler (KL) divergence and reconstruction loss which acts as a model regularizer and restricts the latent vectors to match a Gaussian distribution with a mean 0 and standard deviation 1. Our implementation of VAE is inspired from Way, G.P. *et al*. and is as shown in Figure 2.

We have performed feature reduction across these algorithms for 8 latent dimensionalities (k) in the range k=2 to k=200. The dataset was split into 90% training and 10% test sets balanced by the cancer type. Scikit-learn has been used for the implementation of PCA, ICA and NMF, and for VAE, our model is inspired from the implementation by Way, G. P. *et al*.. These features are fed into the classifier models for binary and multiclass classification. Their performance results are discussed in the next section.

3.2 Cancer Classification

After observing a definite trend in the latent space dimensions, we finally leveraged it for cancer classification. It is a straightforward task to develop a binary classifier that predicts in a 'yes' or 'no' format for all the cancer types, however, from the point of view of scaling the work, this is not an efficient method. Hence, we developed a unified pipeline that is able to accurately predict all 33 cancer types from the TCGA data set. We further experimented with subsets of the data as well.

Our proposed approach makes use of the ensemble method to form a combined vector representing the compressed gene data. It has been selected after a detailed experimentation with other possible permutations of the compression and classification approaches. These have been

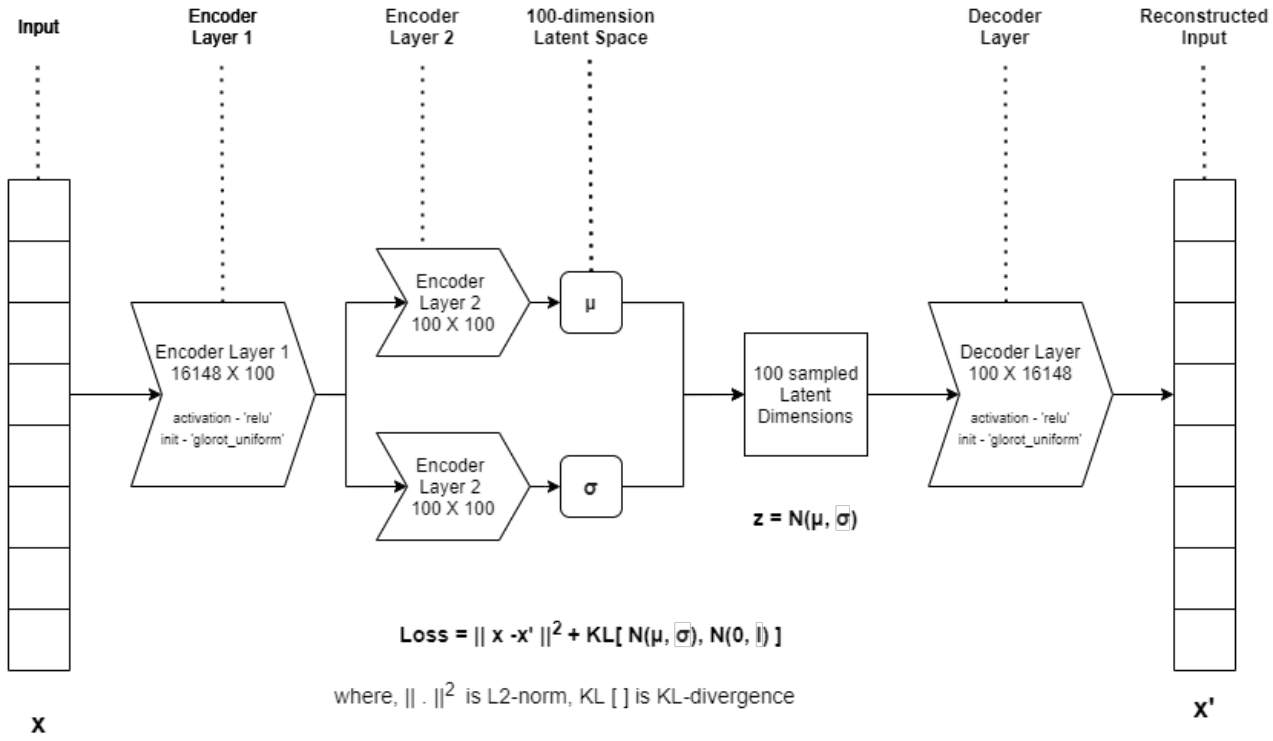

Fig. 2. Our proposed Variational Autoencoder (VAE) Model

discussed later in detail. In our approach these set of compressed features are passed through a single hidden layer neural network that outputs the cancer type. A visual representation of the pipeline can be seen below.

4 Discussion

We report specific and commonly applied performance metrics across all algorithms and latent dimensionalities. The dataset samples were normalized into the range $[0, 1]$ before performing any operation. This range was selected as it is compatible with all our algorithms. The training set was used to train each compression algorithm, that was evaluated on the testing set.

Features	Accuracy	F1-score	Comp. time (s)	K-dimension
Full Data	0.95	0.87	55.06	16137
PCA	0.29	0.17	2.82	50
NMF	0.33	0.19	3.06	50
ICA	0.51	0.21	2.78	50
VAE	0.91	0.89	2.91	100
Ensemble	0.90	0.88	3.64	250

Table 1. Logistic Regression Performance for 33-class Classification

Reconstruction cost is the difference between the input and the reconstructed output. In this case, we have taken the mean square error. This metric indicates the ability of compression models to capture fundamental signals in latent space features that generalizes the original input data. We observe lower reconstruction costs in all the models as the latent dimensionalities increase. As seen in Figure 5 the plots for PCA, and ICA are identical as they are essentially rotations of one another. So, the highest variability exists in the low latent dimensionalities. The next metric is the Pearson correlation between the sample's gene expression input and reconstructed output. It is the ratio of covariance between two standard variables and their standard deviation. This metric indicates the ability of models to capture specific information about sample composition across latent dimensionalities. An overall increasing trend was seen in the median correlation value and decreasing variance as the latent dimensionalities increased. In most of the cases noticed in Figure 5, a gradual increase was seen in the sample correlation values as k increased like in Breast-Invasive Carcinoma (BRCA) or Thyroid Cancer(THCA). However, in some correlation plots, a steep gain is seen with a single increase in the value of k like in Low-Grade Gliomas(LGG) or Diffuse Large B-Cell Lymphoma(DLBC).

Beyond the compression metrics analysis, we analyzed our models based on their classification accuracies. A binary logistic regression classifier was trained using the compression features from each individual algorithm for a one v/s all comparison. In all models, an increase in accuracy was reported with an increase in k, and then the accuracies saturated. For PCA, ICA, and NMF the accuracies saturated for most cancer types at 50 features. For VAE the saturation was seen with 100 features. The accuracies reported were as good as with raw features or even better than them in some cases.

With a multi-class logistic regression classifier for 33 cancer types, the accuracies reported for PCA, ICA, NMF, and VAE were 28.9%, 33.2%, 50.9%, and 90.7% respectively. When this model was trained with an ensemble of all the 250 features, the model accuracy reported was 90.1%. There was still a difference in the accuracy of 3% in this model with the original raw feature space as input. To improve the performance, a deep neural network classifier was employed for these feature sets.

Using this observation, we went ahead with trying the classification task with different deep learning networks as can be seen from Figure 4. The very first experiment was conducted using the vanilla neural network which had the input as the raw gene expression sequence data, a hidden layer of 8000 dimensions and the 33 logits at output. We made use of the

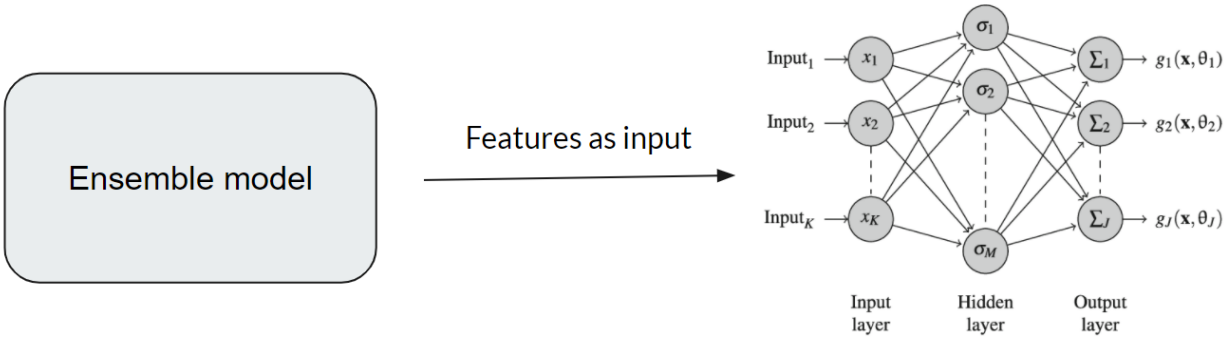

Fig. 3. Neural Network Classifier

Methods & classes	Compression	Classification	Train / Test split	Training time (sec/epoch)	Test Accuracy
Deep learning, 33	None	NN (16147, 8000, 33)	9954 / 1106	2.5	91.32 %
VAE + Deep learning, 33	VAE	NN (100, 50, 33)	9954 / 1106	**0.24**	92.41 %
Our approach, 33	Ensemble	NN (250 ,125, 33)	9954 / 1106	0.31	**94.32 %**
Our approach, 12	Ensemble	NN (250, 125, 12)	6441 / 716	0.3	**97.2 %**

Fig. 4. Neural Network classification accuracy comparison

Cross Entropy Loss function. This model was able to achieve a 91.3% accuracy over the test set. Although, this achieves a higher accuracy as compared to the logistic regression method, it is still not practical to implement it in real life scenarios. Also, it took a training time of around 15 min to train on 9k samples. From a scaling perspective this is a huge drawback of the network. Moving ahead, we experimented with the compressed feature sets obtained from the various algorithms and have highlighted the ones with promising results in the experiments table below.

While using VAE as the compression algorithm followed by a vanilla neural network with the specifications shown in the table, we were able to predict the cancer types with 92.41% accuracy on the test set. This method had a slight improvement in terms of the accuracy as compared to the Deep Learning network, however, we achieved a significant reduction in the training time to 30 seconds for the entire train set. In the next phase, we tried out the ensemble method which surprisingly presented the best results at 94.32% accuracy for the test set.

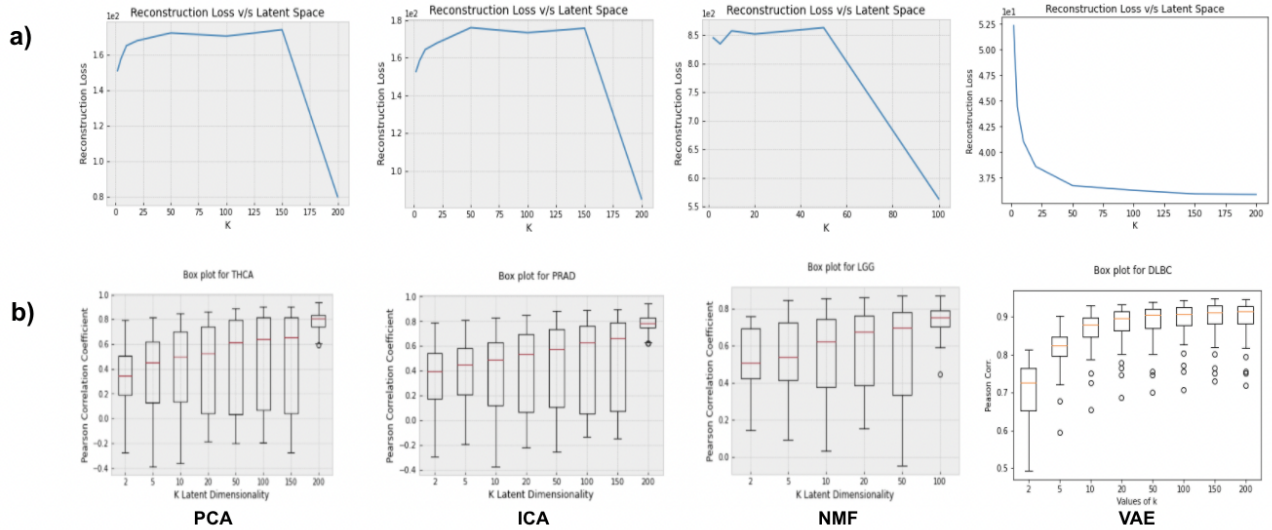

Fig. 5. a. Reconstruction cost and b. Pearson correlation plots for PCA, ICA, NMF, and VAE.

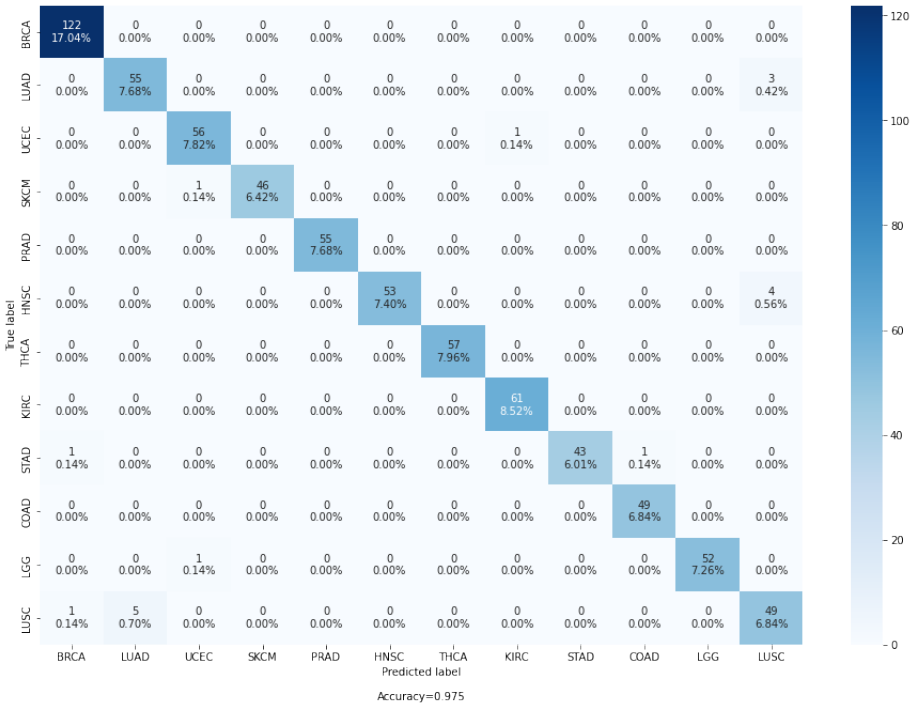

Fig. 6. Confusion Matrix for 12-class classification

After analyzing the dataset class distribution, we observed that it has a high class imbalance amongst the 33 classes. In order to attain an equal distribution, we decided to form a subset of the data and kept 12 classes which have approximately equal distribution. Our model achieved a 97.5% accuracy on this subset of the data which translates to very few mispredictions across, hence, making it a practical model to implement in real life scenarios. Refer Figure 6 for the results of this experiment.

5 Conclusion

We concluded from our experiments that biological representations are enhanced for cancer type classifications when gene expression data is compressed using linear, and non-linear algorithms into many latent space dimensionalities. An ensemble of these features classified using a deep neural network for 12 classes gives the best accuracy 97.5 %, higher than the accuracy obtained in the original genomic space, with an exponentially lesser computational time. On lowering the class count in our model, we reach an almost perfect accuracy score. VAE performs well in all metrics across experiments, however, the ensemble model outperforms it for the subtask of cancer classification using neural networks.

As machine learning continues to be applied to derive insight from biomedical data sets, researchers should shift focus away from optimizing a single model based on certain mathematical heuristics, and instead towards learning good and reproducible biological representations that generalize to alternative data sets regardless of compression algorithm and latent dimensionality. Subtle patterns in input signals can be identified in this approach which aides the task of cancer classification.

Acknowledgements

We would like to extend our sincere gratitude to Prof. Sriram Sankararaman and Teaching Assistant, Boyang Fu for guiding us throughout the course of this project and giving us valuable feedback to make our project successful.

References

Way, G.P., Zietz, M., Rubinetti, V. et al. (2020) Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations., *Genome Biol*, **21**, 109.

Jolliffe IT, Cadima J. (2016) Principal component analysis: a review and recent developments, *Philos Transact A Math Phys Eng Sci.*, **374**, 20150202.

Kairov U, Cantini L, Greco A, Molkenov A, Czerwinska U, Barillot E, et al (2017) Determining the optimal number of independent components for reproducible transcriptomic data analysis, *BMC Genomics.*, **18**, 712.

Frigyesi, A. & Höglund, M (2008) Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes, *Cancer Inf.*, **6**, 275–292.

Way, G. P., & Greene, C. S. (2018). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders, *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, **23**, 80-91.