Unraveling the isoform world: a VAE approach with a FFNN (Project 29)

Laura Figueiredo Tor (s222797), Ana Pastor Mediavilla (s222761),
Carlos de Santiago León (s222766), Anu Oswall (s222498)

02456 Deep learning
Project 29

1. Introduction

- Analysis of **isoform expression** is essential as isoforms play key roles in biological processes and can be used as therapeutic targets or biomarkers[1].
- Predicting isoform expression from transcriptomics data using Deep Learning models would allow us to understand the impact of these isoforms without needing to do further analyses, shedding light on the effects of genetic variants and protein functions on specific tissues.
- A **Latent-feature discriminative model** using a Variational Autoencoder (VAE) would enable to map the input data to a reduced latent representation, therefore capturing the key features. It would provide a low-dimensional representation of the data, which is of special interest in high-dimensionality datasets as the transcriptomics datasets[2].

2. Data

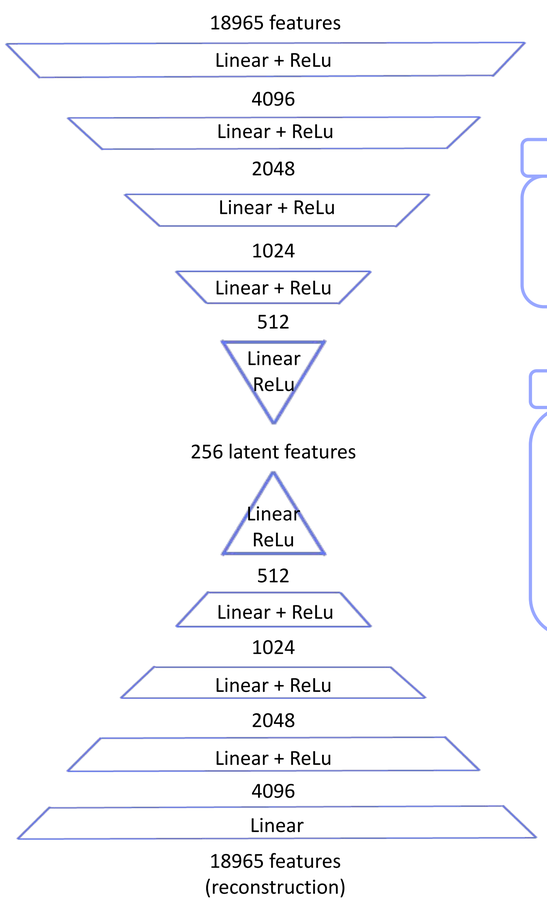
Gene-expression data, $Log_2(TPM+1)$, from different human tissues:

- **archs4 gene-expression dataset [167883, 18965]:** big un-labeled gene expression data used for VAE's training.
- **gtex gene-expression dataset [17355, 18965]:** small paired data containing gene-expression (X for predictive model).
- **gtex isoform-expression dataset [17355, 156958]:** small paired data containing isoform-expression (y for predictive model).
- *gtex gene isoform annotation file:* gene-isoform relationship for gtex dataset.
- *gtex tissue annotation:* tissue types (54) for each sample.

3. Methods

Latent-feature discriminative model (M1) approach[2]

VAE



Distributions

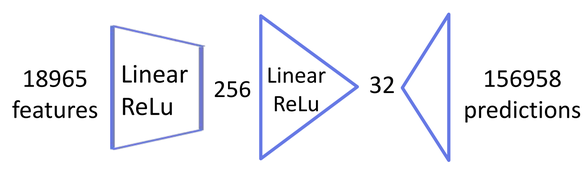
Prior: Gaussian
Posterior: Gaussian
Observation model: Log-normal (unit-variance)

Validation

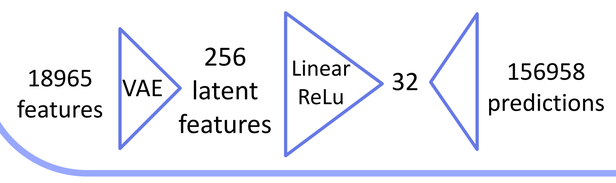
$$\mathcal{L}(\mathbf{x}) := \overbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}^{\text{(a) Reconstruction Error}} - \beta \cdot \overbrace{\mathcal{D}_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \mid p(\mathbf{z})\right)}^{\text{(b) Regularization}}$$

ANN

FFNN for Non-Encoded data (2 hidden layers):



FFNN for Encoded data (1 hidden layer):



Train-test split

With stratification
- Training: 13885 samples
- Validation: 3471 samples

Validation

Stratified cross-validation

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

4. Results

Variational AutoEncoder

5 layers and LogNormal VAE observation model

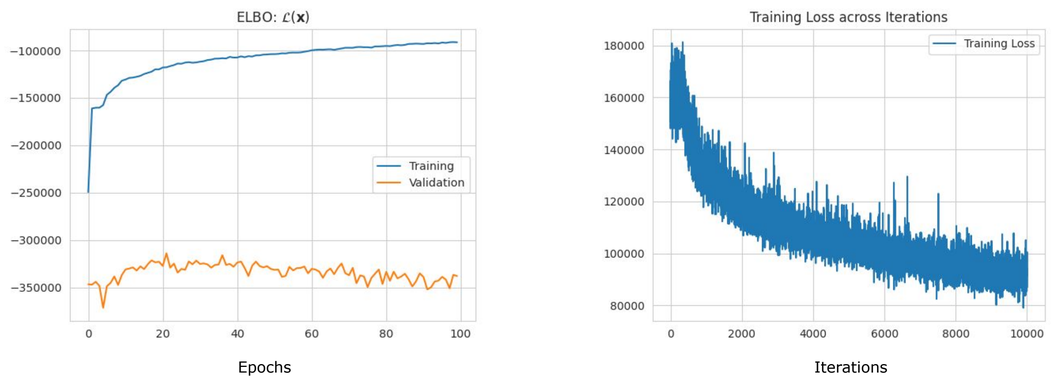


Figure 1. Evidence Lower Bound (ELBO) and training loss across all of the epochs and iterations, respectively. On the left, the ELBO values across the different epochs, on the right, the loss values across the 100 epochs for the 5 layer VAE-model with a LogNormal observation model.

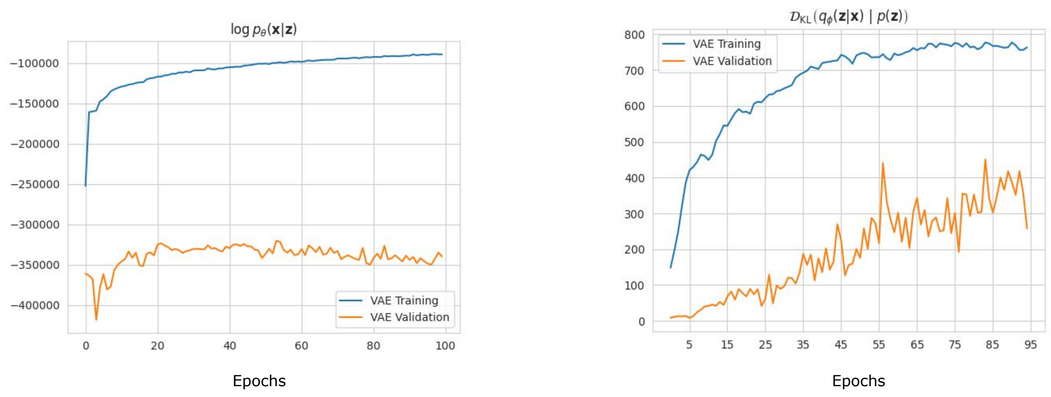


Figure 2. Negative likelihood (NLL) and KL-divergence evolution across all of the batches for the training of the VAE. On the left, the NLL values across the different epochs, on the right, the KL-term values for the 95 last epochs for the 5 layer VAE-model with a LogNormal observation model.

Feed Forward Neural Network

Table 1. Mean Squared Error (MSE) performance from the FFNN trained for 5 stratified folds and 50 training epochs. "Non Encoded Data" refers to the raw gtex dataset without feature reduction, "VAE Encoded Data" refers to the gtex dataset after reducing the features space with the encoder of the VAE trained with the archs4 dataset, and "Baseline" refers to the simplest model where the prediction is based on the mean value for each isoform (in each validation batch). The p-values, obtained through t-tests, assess the statistical significance of the differences in performance between the models.

Cross Validation Fold	MSE VAE Encoded Data (1 hidden layer)	MSE Non Encoded Data (2 hidden layers)	MSE baseline	p-value VAE vs Non-VAE	p-value VAE vs baseline	p-value VAE vs baseline
1	0.31993	0.47646	0.48407	8.375 e-46	1.015 e-50	0.042
2	0.31285	0.47814	0.48469	5.167 e-44	1.841 e-59	0.113
3	0.32756	0.47812	0.48801	4.132 e-40	1.857 e-57	0.027
4	0.32657	0.47506	0.48632	1.002 e-45	2.223 e-46	0.006
5	0.31940	0.47538	0.48334	7.449 e-41	2.04 e-52	0.078

5. Discussion and conclusions

- Variational AutoEncoders could be useful for feature reduction: our VAE implementation is performing accurately by observing a minimization of the ELBO.
- Prediction of all isoforms seems to be efficient with feature reduction: a high-feature dimension seems to be inaccurate to predict all the isoform-expressions.
  - Most of the isoforms have low expression, which makes it difficult to have a FFNN that performs better than the baseline.
- It would be interesting to analyse how the network could generalise to unseen tissues by testing it on a independent test set: this could be implemented with our dataset while keeping the stratification structure.
- VAE performance could improve if training with all the available dataset was possible.
- Deeper analysis for refining the network hyperparameters should be carried out.

References

1. Safikhani, Z., Smirnov, P., Thu, K.L. et al. Gene isoforms as expression-based biomarkers predictive of drug response in vitro. Nat Commun 8, 1126 (2017). https://doi.org/10.1038/s41467-017-01153-8.
2. Kingma, D. P., Mohamed, S., Jimenez Rezende, D., & Welling, M. Semi-supervised learning with deep generative models. Advances in neural information processing systems, 27 (2014). https://doi.org/10.48550/arXiv.1406.5298.