Vilniaus universitetas

Doktoranto Karolio Šablausko ataskaita už 2023/2024 mokslo metų pirmąjį pusmetį

Darbo vadovas: prof. Audronė Jakaitienė

Vilniaus universitetas

Disertacijos pavadinimas: Characterization of genetic changes using deep neural networks

Doktorantūros pradžios ir pabaigos metai: 2022 – 2026

Studijų metai: 2.

Vilniaus universitetas

1 lentelė: Doktorantūros studijų planas

Studijų metai	Egzaminai	
	Planas	Įvykdyta
I (2022/2023)	1	1
II (2023/2024)	2	1 (+ 1 numatomas 2024-06 mėn)
III (2024/2025)	1	0
IV (2025/2026)	0	0
Iš viso:	4	0

Studijų metai	Dalyvavimas konferencijose				Publikacijos					
	Tarptautinėse		Nacionalinėse		Su citavimo rodikliu			Be citavimo rodiklio		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė	Planas	Įvykdyta	Būklė
I (2022/ 2023)	0	0	0	0	0	0		0	0	
II (2023/ 2024)	1	0	0	0	0	0		0	0	
III (2024/ 2025)	1	0	0	0	1	0		0	0	
IV (2025/ 2026)	1	0	0	0	1	0		0	0	
Iš viso:	3	0	0	0	2	0		0	0	

1 Thesis design

1.1 Thesis aim

- To contribute to the advancement of deep learning techniques for the analysis of next generation sequencing data, with a focus on single cell RNA sequencing (scRNA-seq) data.

1.2 Thesis objectives

- **Data preprocessing and feature engineering**: create efficient data preprocessing pipeline suited for scRNA-seq data, including normalization and batch effect correction to prepare the data for further analysis.
- **Deep differential expression analysis**: develop and implement deep learning-based approach for identifying differentially expressed genes and pathways.
- **Evaluation and benchmarking**: conduct extensive benchmarking and cross-validation experiments to assess the performance and generalizability of deep learning models, comparing them to traditional methods.
- **Biological case study**: apply said techniques in the interpretation of biological data gathered during a biomedical study.

DATA STRUCTURE

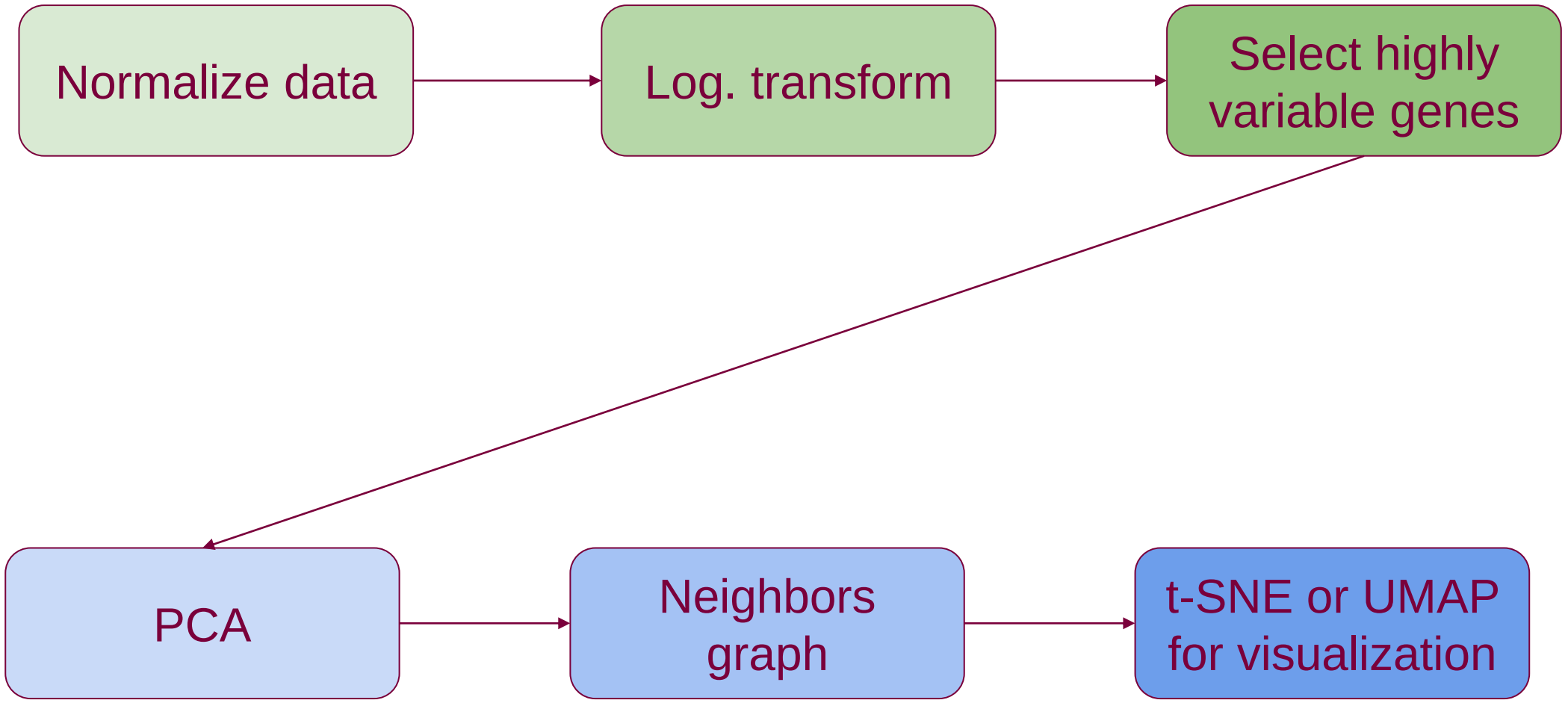
Vilniaus universitetas

	cell_0	**cell_1**	**…**	**cell_N**
gene_0	12	0		180
gene_1	0	20		0
…				
gene_M	0	15		0

matrix **M x N - Count matrix**
M = 37 000
N = 200 000

Vilniaus universitetas

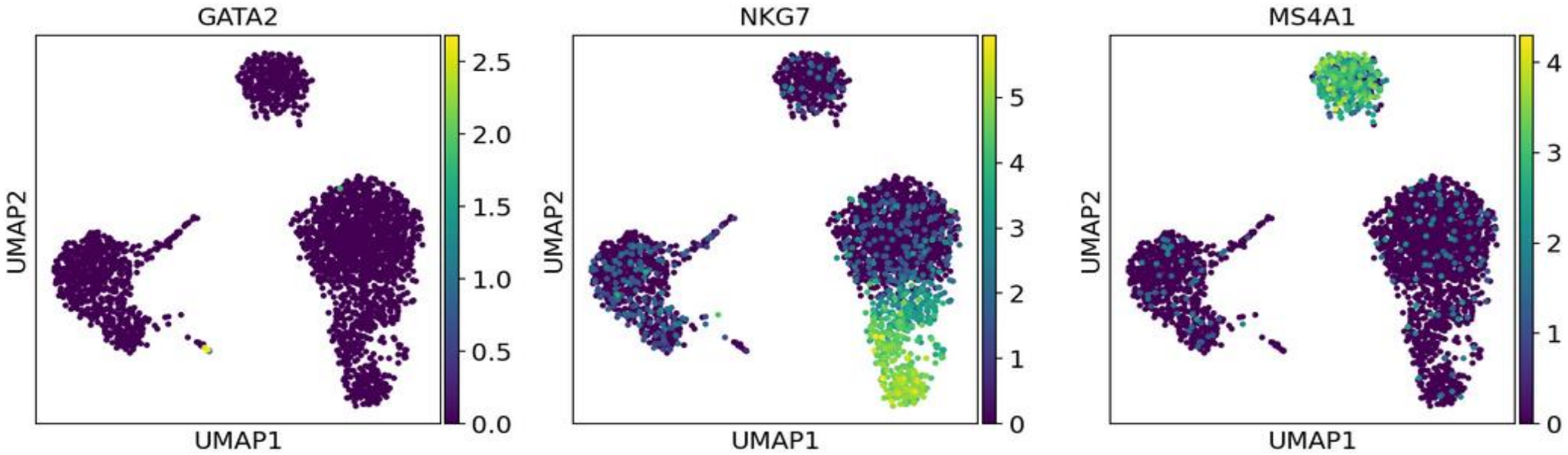
COUNT MATRIX PROCESSING



DIMENSIONALITY REDUCTION

Vilniaus universitetas

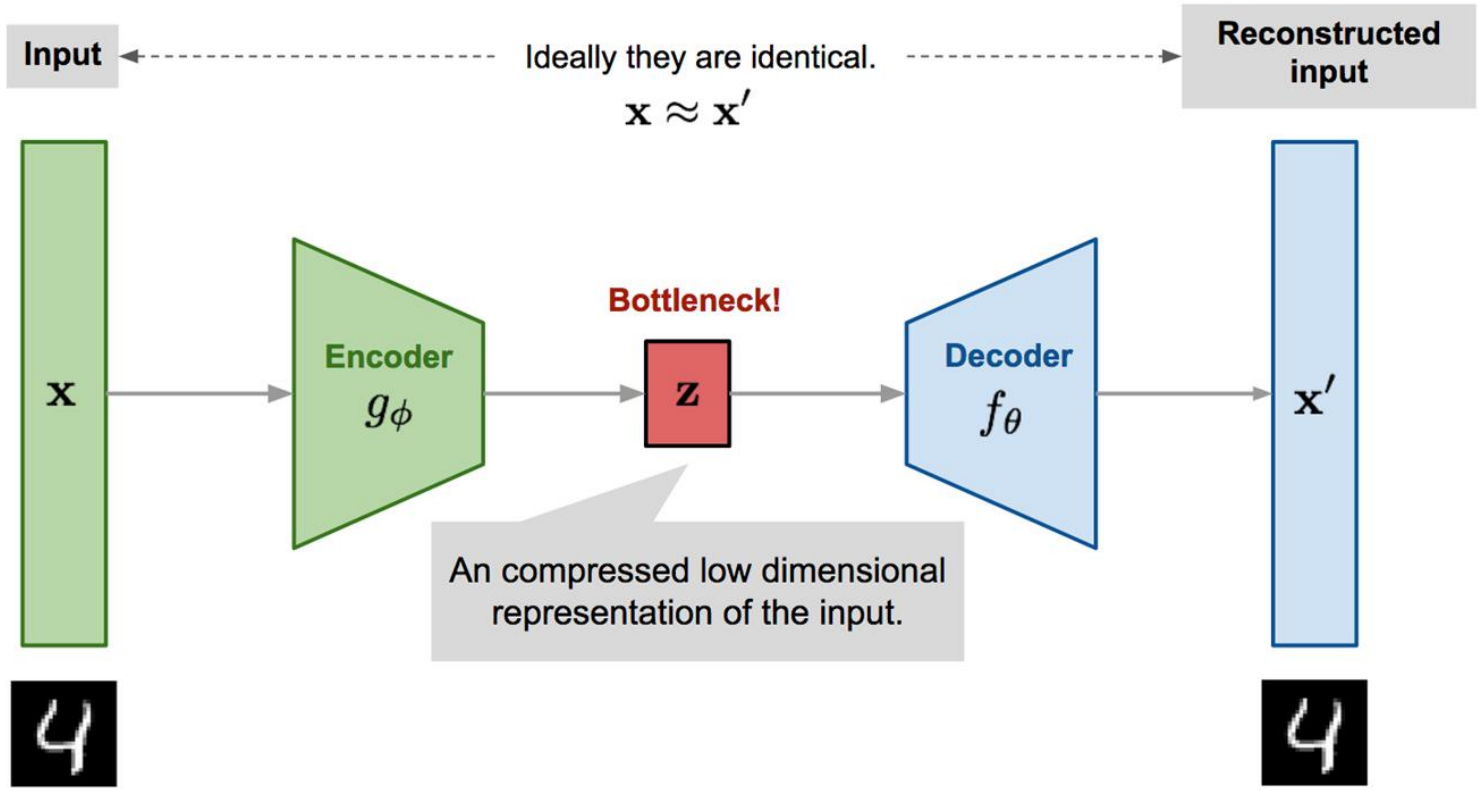
UMAP RESULT



2000 dimensions → 2 dimensions

Vilniaus universitetas

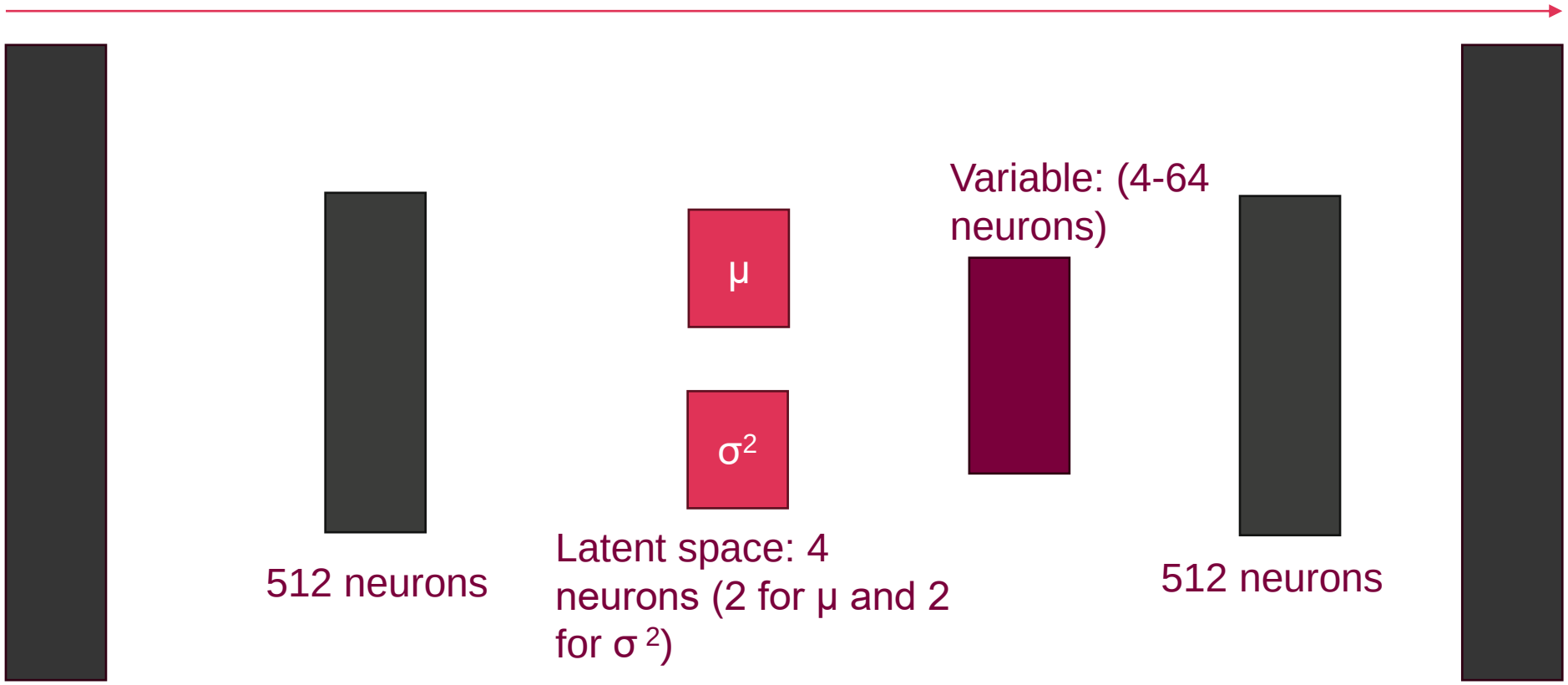
Variational autoencoder



https://lilianweng.github.io/posts/2018-08-12-vae/

VARIATIONAL AUTOENCODER ARCHITECTURE

Vilniaus universitetas



Input -> cell
Vector with size = gene count

Output -> reconstructed cell
Vector with size = gene count

Vilniaus universitetas

SINGLE CELL PIPELINE USING VAE