Doktoranto Karolio Šablausko ataskaita už 2024/2025 mokslo metų pirmą pusmetį

- **Disertacijos pavadinimas:** Genetinių pokyčių charakterizavimas naudojant giliojo mokymo neuroninius tinklus (angl. Characterization of genetic changes using deep neural networks)
- **Darbo vadovas:** prof. Audronė Jakaitienė
- **Doktorantūros pradžios ir pabaigos metai:** 2022 – 2027 (akademinės atostogos 2024-10-01 – 2025-01-31)
- **Studijų metai**: 3.

Studijų metai	Egzaminai	
	Planas	Įvykdyta
I (2022/2023)	1	1
II (2023/2024)	2	2
Akademinės atostogos 2024-10-01 – 2025-01-31		
III (2025/2026)	1	0
IV (2026/2027)	0	0
Iš viso:	4	3

Išlaikytas egzaminas dalykui "Didžiųjų duomenų analitika".

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	0	0
II (2023/2024)	1	0	0	0	0	0	0	0	0	0
Akademinės atostogos 2024-10-01 – 2025-01-31										
III (2025/2026)	1	0	0	0	1	0	1 pateikta publikacija	0	0	-
IV (2026/2027)	1	0	0	0	1	0		0	0	-
Iš viso:	3	0	0	0	2	0		0	0	-

High-throughput single cell -omics using semi-permeable capsules

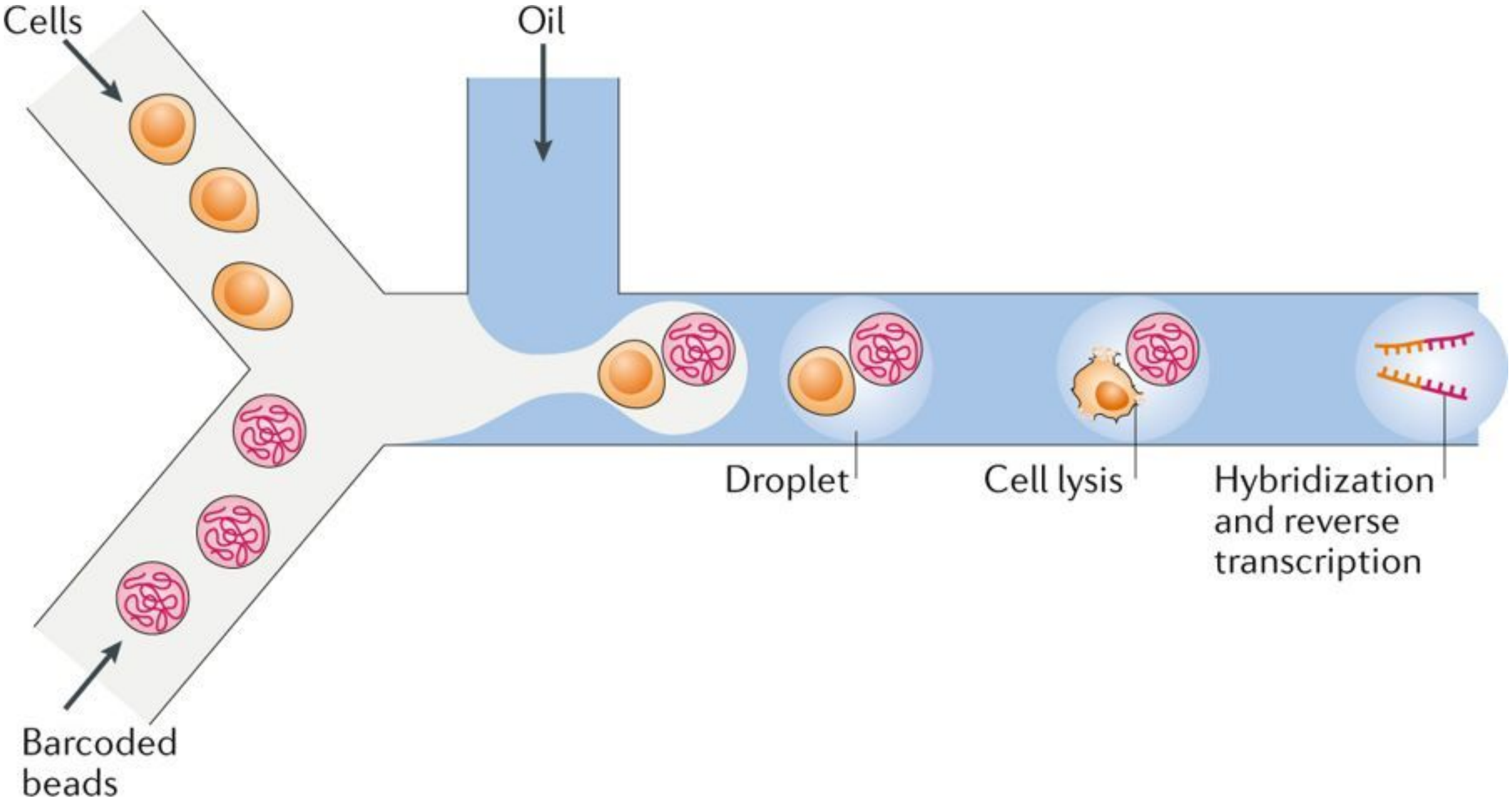
Denis Baronas[1], Justina Zvirblyte[1], Simonas Norvaisis[1], Greta Leonaviciene[1], Karolis Goda[1], Vincenta Mikulenaite[1], Vytautas Kaseta[3], Karolis Sablauskas[2,4], Laimonas Griskevicius[2], Simonas Juzenas[1] and Linas Mazutis[1,5,*]

[1] Institute of Biotechnology, Life Sciences Center, Vilnius University, Vilnius, Lithuania
[2] Hematology, Oncology and Transfusion Medicine Center, National Cancer Center, Vilnius University Hospital Santaros Clinics, Vilnius, Lithuania
[3] Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, Lithuania
[4] State Research Institute Centre for Innovative Medicine, Department of Stem Cell Biology, Vilnius, Lithuania
[5] Department of Molecular Biology, Umea University, Sweden
[#] Correspondence

Abstract

Biological systems are inherently complex and heterogeneous. Deciphering this complexity increasingly relies on high-throughput analytical methods and tools that efficiently probe the cellular phenotype and genotype. While recent advancements have enabled various single-cell -omics assays, their broader applications are inherently limited by the challenge of efficiently conducting multi-step biochemical assays while retaining various biological analytes. Extending on our previous work (*1*) here we present a versatile technology based on semi-permeable capsules (SPCs), tailored for a variety of high-throughput nucleic acid assays, including digital PCR, genome sequencing, single-cell RNA-sequencing (scRNA-Seq) and FACS-based isolation of individual transcriptomes based on nucleic acid marker of interest. Being biocompatible, the SPCs support single-cell cultivation and clonal expansion over long periods of time – a fundamental limitation of droplet microfluidics systems. Using SPCs we perform scRNA-Seq on white blood cells from patients with hematopoietic disorders and demonstrate that capsule-based sequencing approach (CapSeq) offers superior transcript capture, even for the most challenging cell types. By applying CapSeq on acute myeloid leukemia (AML) samples, we uncover notable changes in transcriptomes of mature granulocytes and monocytes associated with blast and progenitor cell phenotypes. Accurate representation of the entirety of the cellular heterogeneity of clinical samples, driving new insights into the malfunctioning of the innate immune system, and ability to clonally expand individual cells over long periods of time, positions SPC technology as customizable, highly sensitive and broadly applicable tool for easy-to-use, scalable single-cell -omics applications.

- Išlaikytas 1 egzaminas (3/4) - dar 1 numatomas šį semestrą
- Pateiktas manuskriptas publikacijai su GMC tyrėjais žurnalui *Nature* (IF 50.5). Autorių sąraše 8/11.
- Ruošiama publikacija pirmuoju autorium į *Bioinformatics* (IF 6.9) arba *Nucleic Acid Research* (16.7).
- Konferencijos - 1 abstraktas buvo pateiktas, bet nebuvo priimtas. Numatoma teikti 2 šį semestrą.
- Stažuotė numatoma 2025-05 mėn.



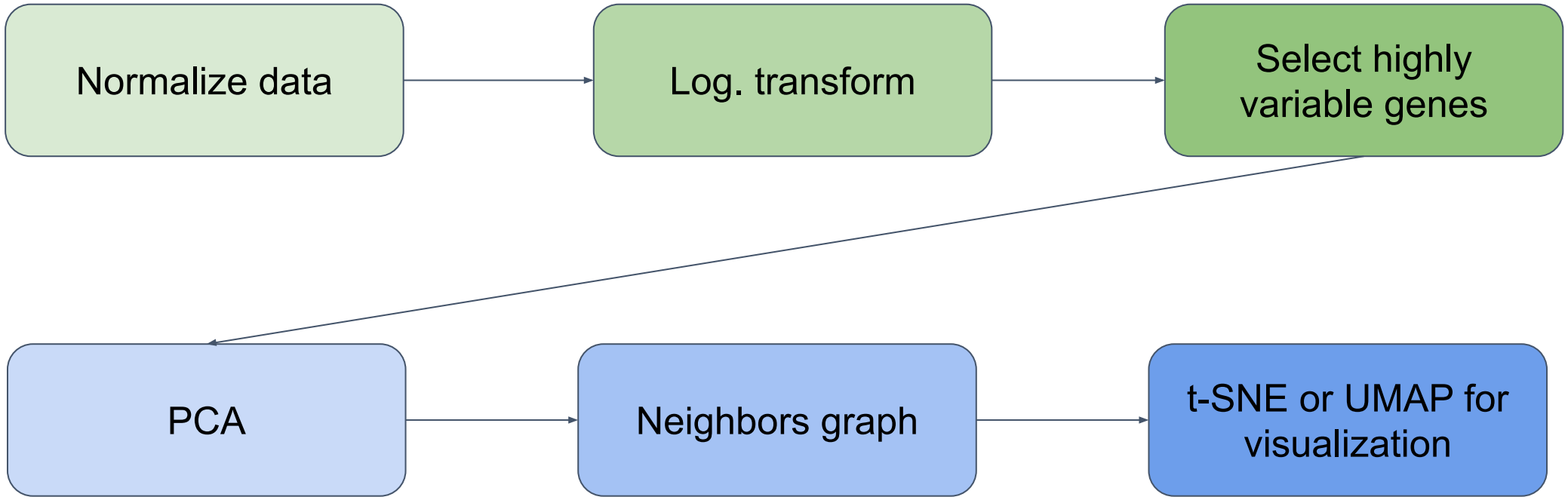
https://www.rna-seqblog.com/single-cell-rna-sequencing-for-the-study-of-development-physiology-and-disease/

DATA STRUCTURE

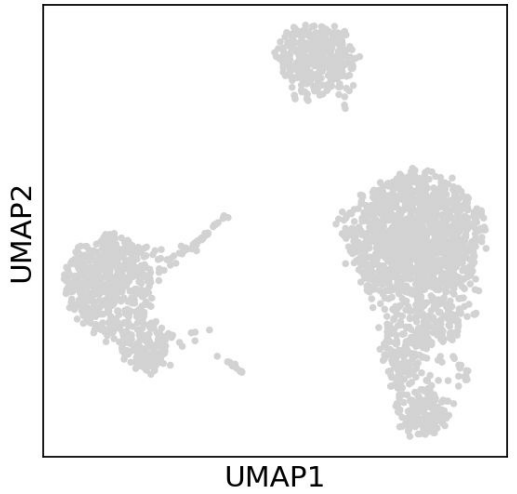
	cell_0	**cell_1**	**…**	**cell_N**
gene_0	12	0		180
gene_1	0	20		0
…				
gene_M	0	15		0

matrix **M x N - Count matrix**
M = 2*10e3 - 3*10e4
N = 2*10e3 - 2*10e6

COUNT MATRIX PROCESSING



DIMENSIONALITY REDUCTION

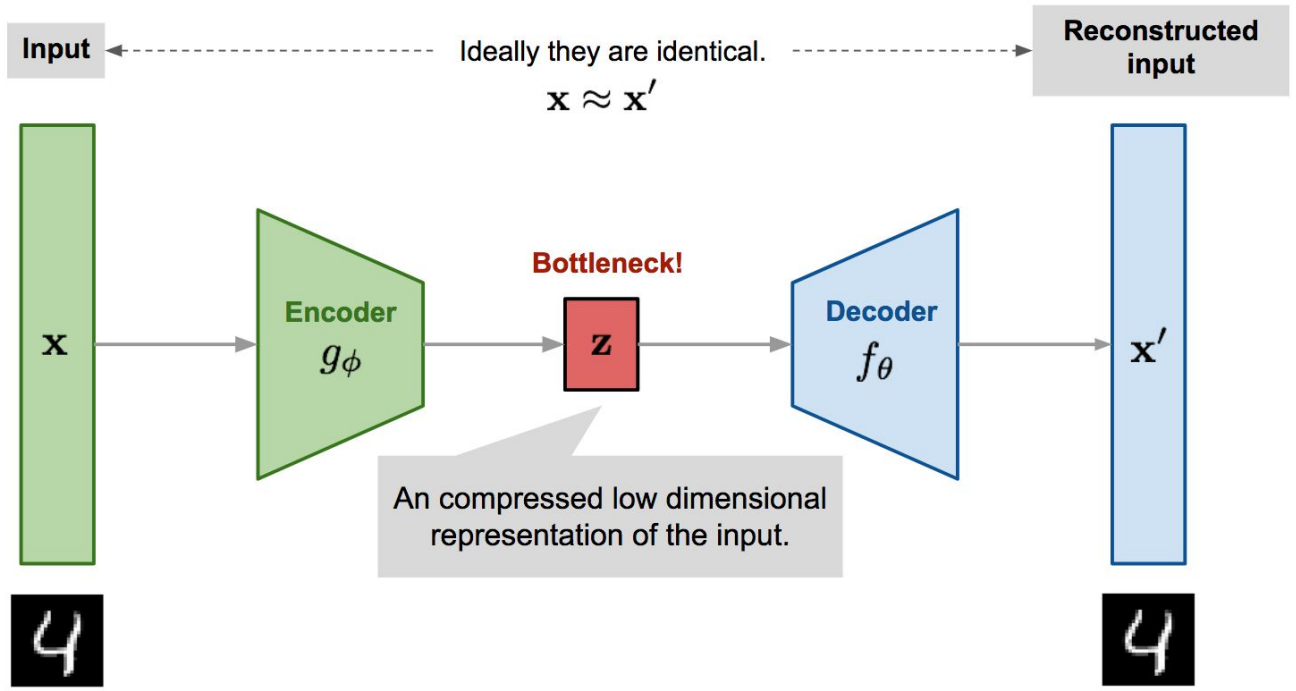


Challenges related to current approach:

- a number of parameters are chosen in an **arbitrary** way
- each change requires **regenerating** projection

Step	Change in the number of dimensions	Reasoning	Is the choice of parameters arbitrary?
Selection of variable genes - biological priors	37,773 →37,528	Mitochondrial and ribosomal genes are non-differentiating so these features can be removed	No - mitochondrial and ribosomal gene sets are well defined.
Selection of variable genes - metric-based approach	37,528 → 3,000	Fano factor (or other variance metric) allows selecting highly variable or highly informative genes	Depends on the specific approach used. In the case of the Fano factor - yes.
PCA	3,000 →93	Selects only the highly informative principal components	No - number of principal components can be chosen using the elbow method
KNN	not applicable	Constructs a graph of *k* nearest neighbors for each cell.	Yes - “rules of thumb” exist, for example *k* should be the minimum number of cells expected for each type.
UMAP	93 →2	Computes the local and global structure of the data.	Yes - requires experimentation to find optimal spacing of cells.

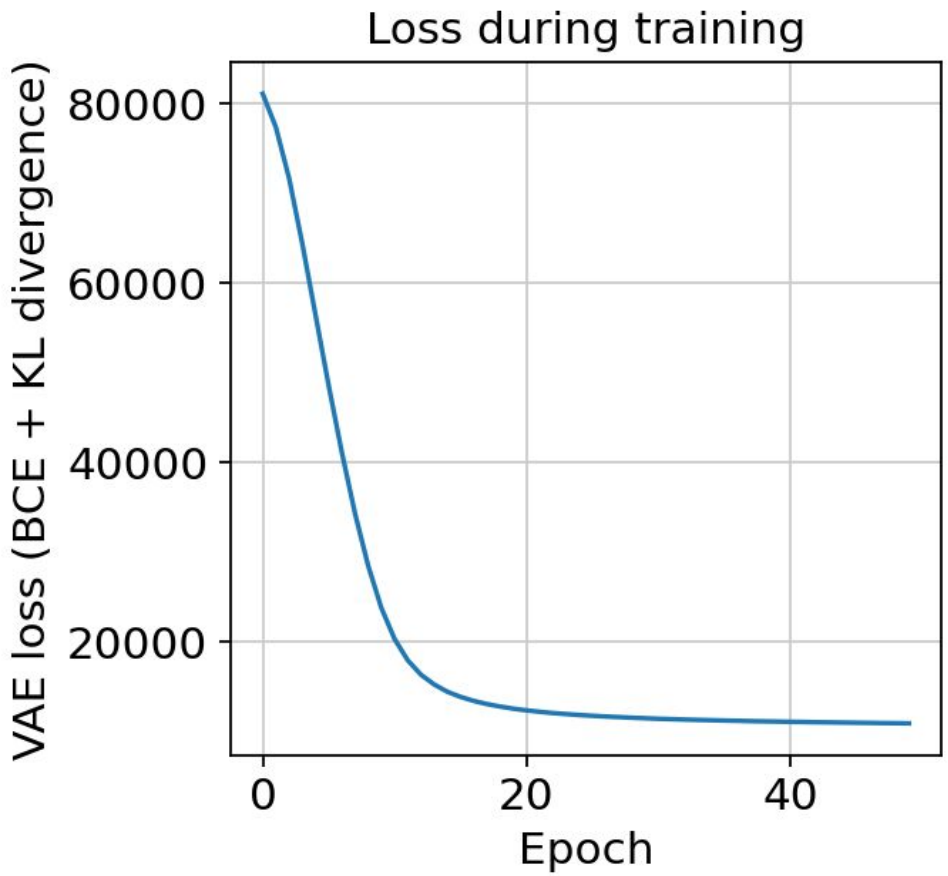
Variational autoencoder



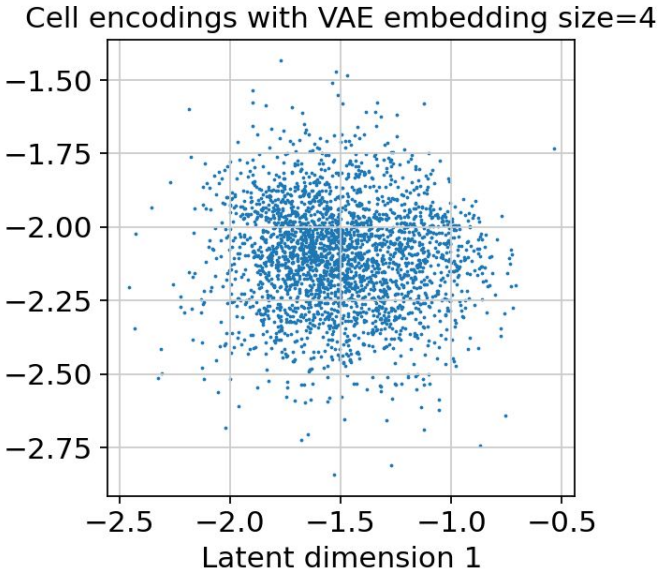
https://lilianweng.github.io/posts/2018-08-12-vae/

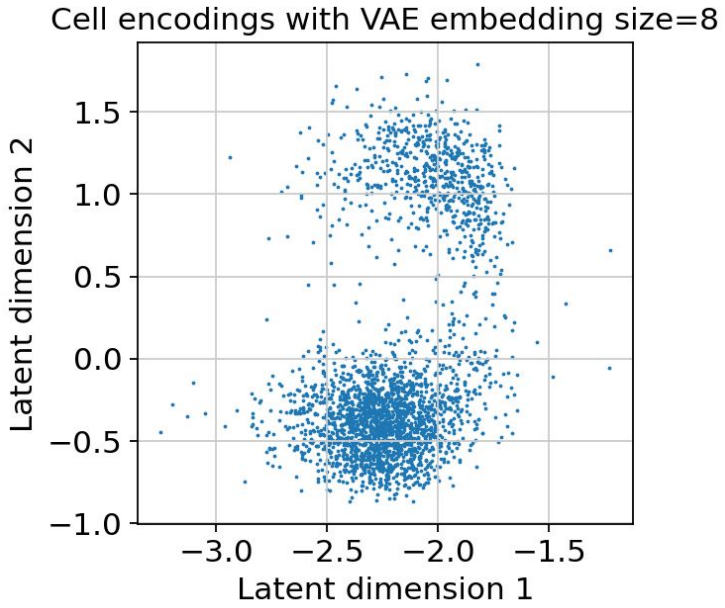
```
:
# Variational Autoencoder model for gene embedding
class VAE(nn.Module):
    def __init__(self, input_size, embedding_size, latent_size):
        super(VAE, self).__init__()
        self.fc1 = nn.Linear(input_size, 512)
        self.fc_mean = nn.Linear(512, latent_size)
        self.fc_logvar = nn.Linear(512, latent_size)
        self.fc2 = nn.Linear(latent_size, embedding_size)
        self.fc3 = nn.Linear(embedding_size, input_size)

    def encode(self, x):
        x = torch.relu(self.fc1(x))
        mean = self.fc_mean(x)
        logvar = self.fc_logvar(x)
        return mean, logvar

    def reparameterize(self, mean, logvar):
        std = torch.exp(0.5 * logvar)
        eps = torch.randn_like(std)
        return mean + eps * std

    def decode(self, z):
        z = torch.relu(self.fc2(z))
        x_hat = torch.sigmoid(self.fc3(z))
        return x_hat

    def forward(self, x):
        mean, logvar = self.encode(x)
        z = self.reparameterize(mean, logvar)
        x_hat = self.decode(z)
        return x_hat, mean, logvar
```

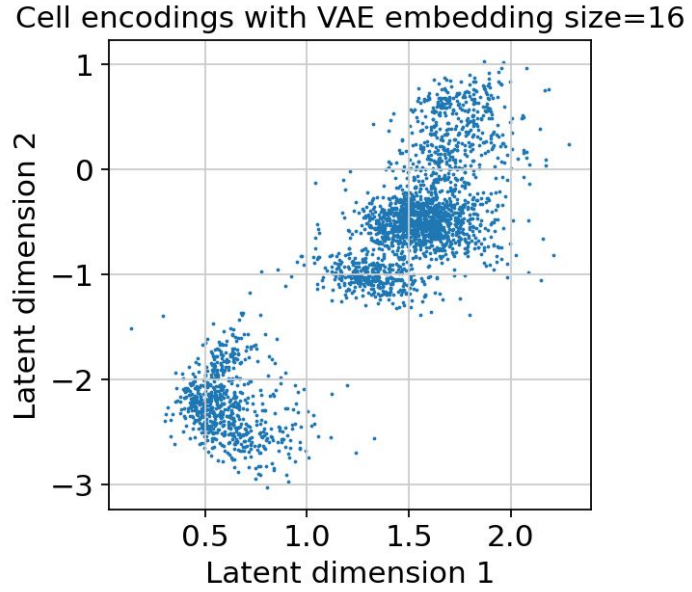
```python
# Define hyperparameters
input_size = count_matrix.shape[1]  # Number of genes in the count matrix
embedding_size = 32  # Embedding dimension
latent_size = 2  # Latent variable dimension
learning_rate = 0.001
num_epochs = 50
batch_size = 64
```

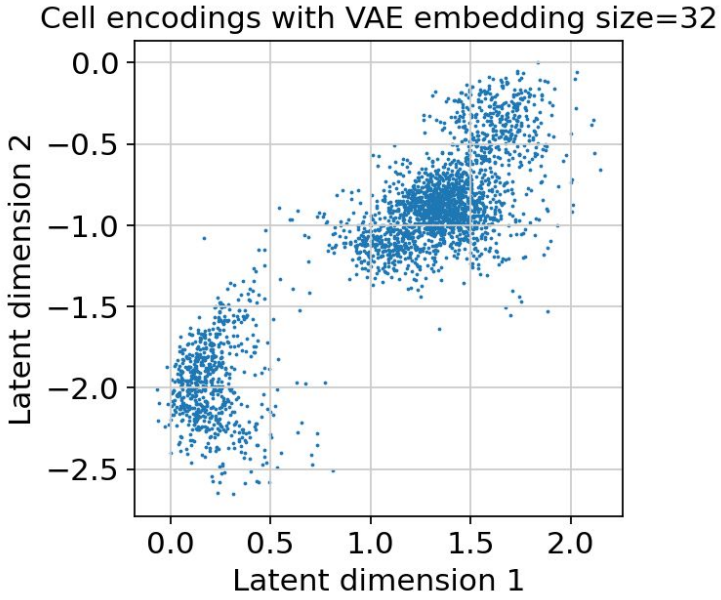


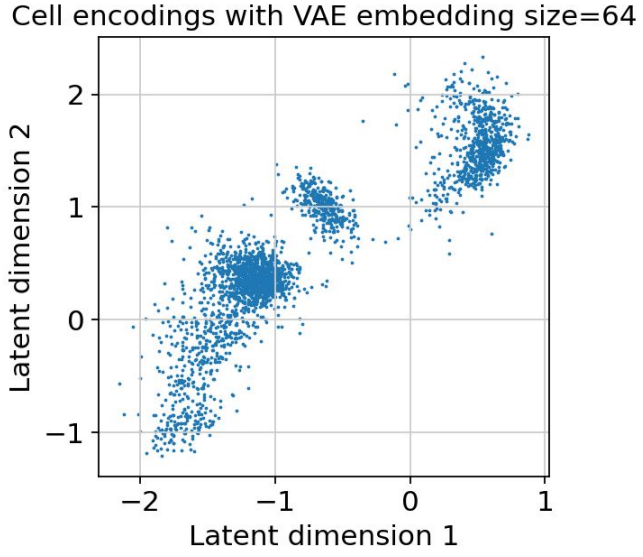
Testing hyperparameter effects on small dataset - 3000 cells



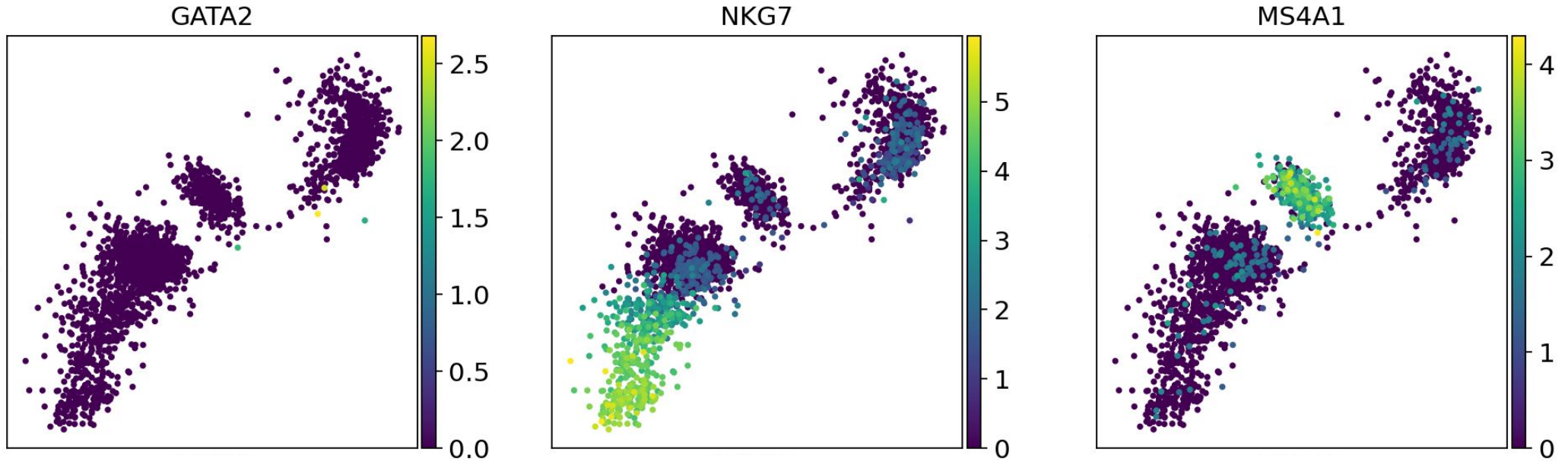




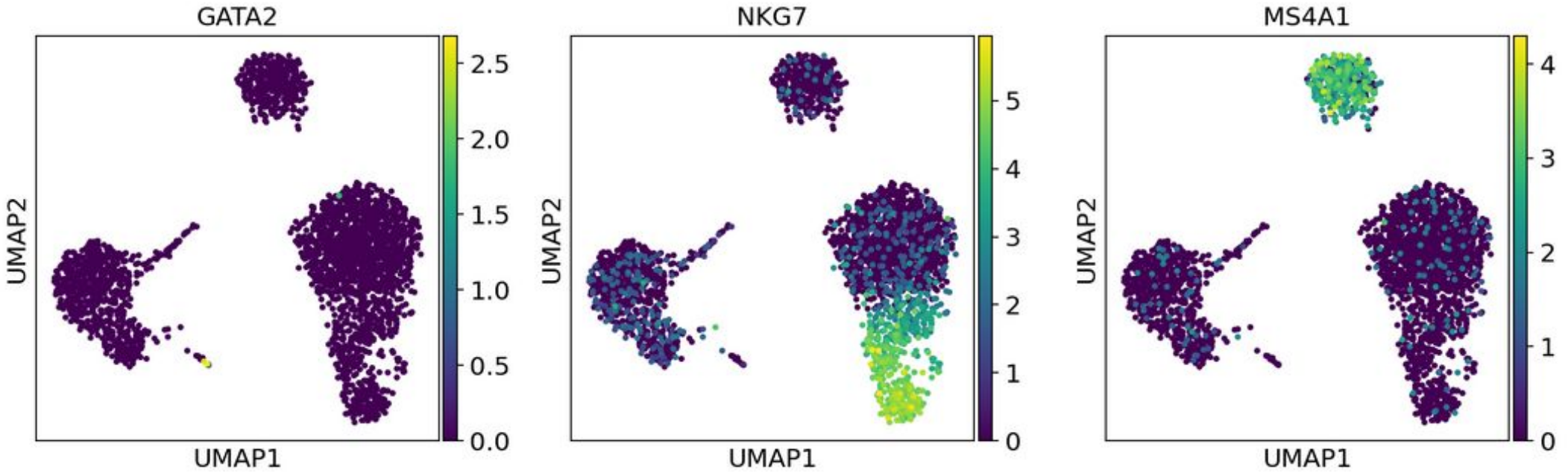




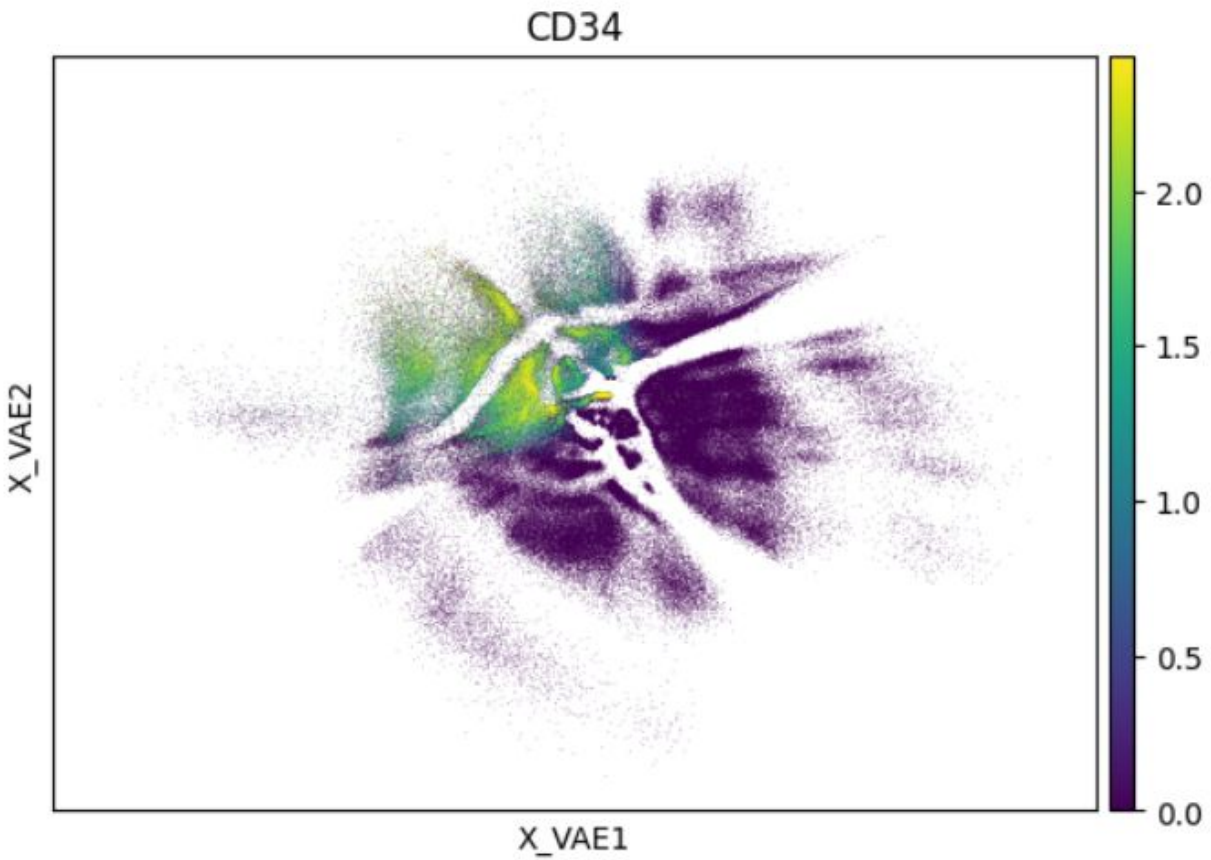
VAE, embedding size = 64 latent space

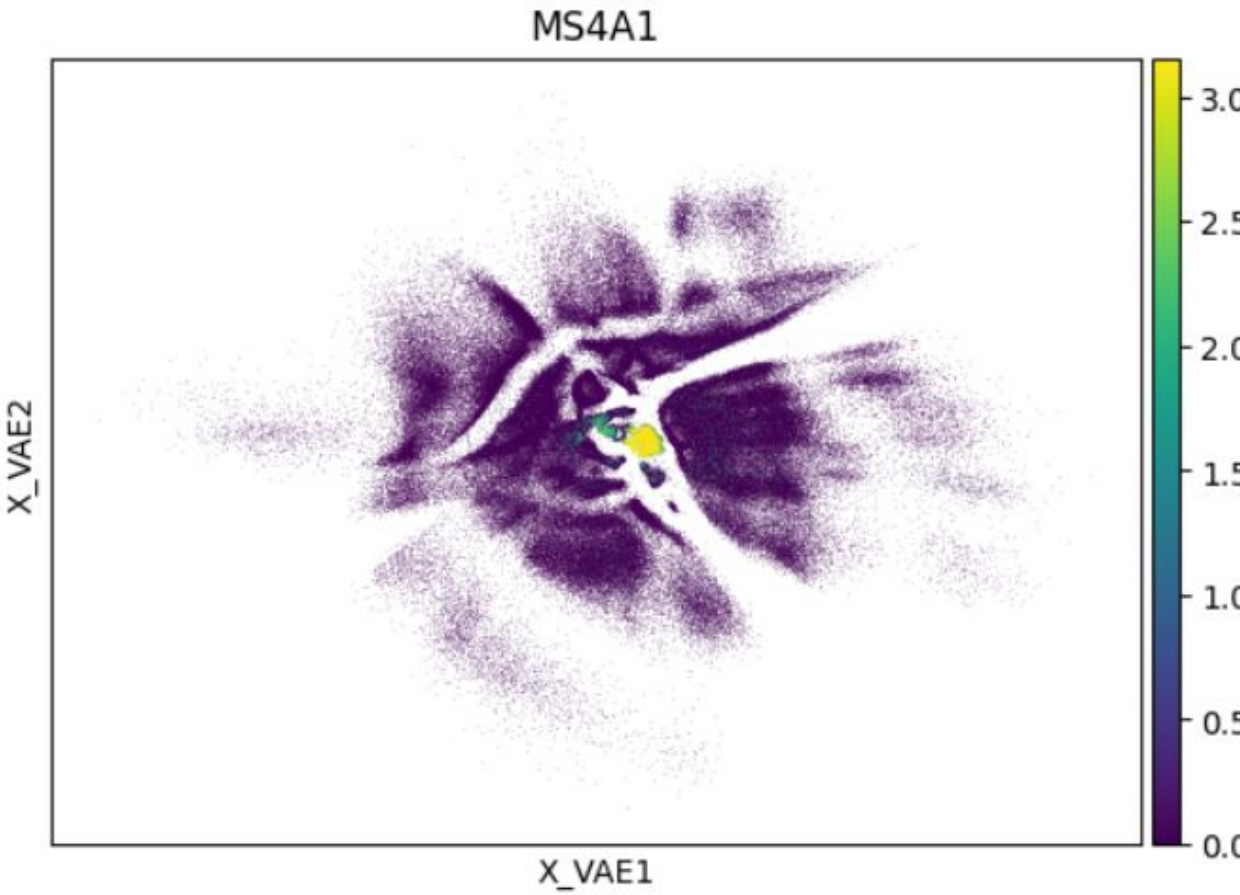


UMAP

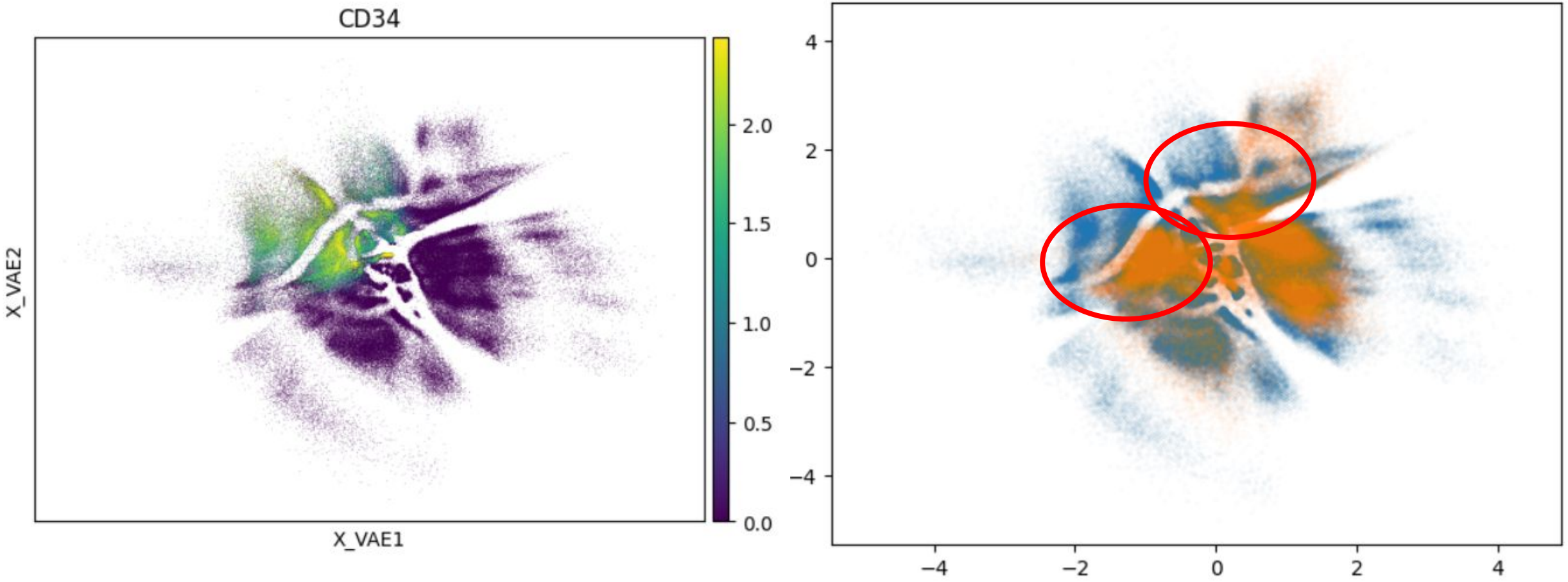


Training VAE on large dataset: 263,159 cells





Projecting 152,053 tumor cells on the trained VAE



2 publikacijos su DMSTI afiliacijomis (doktorantūros tezei nebus naudojamos)

nature medicine IF 58.7

Explore content · About the journal · Publish with us

nature › nature medicine › articles › article

Article | Open access | Published: 17 January 2025

Genomic reanalysis of a pan-European rare-disease resource yields new diagnoses

Steven Laurie, Wouter Steyaert, Elke de Boer, Kiran Polavarapu, Nika Schuermans, Anna K. Sommer, German Demidov, Kornelia Ellwanger, Ida Paramonov, Coline Thomas, Stefan Aretz, Jonathan Baets, Elisa Benetti, Gemma Bullich, Patrick F. Chinnery, Jill Clayton-Smith, Enzo Cohen, Daniel Danis, Jean-Madeleine de Sainte Agathe, Anne-Sophie Denommé-Pichon, Jordi Diaz-Manera, Stephanie Efthymiou, Laurence Faivre, Marcos Fernandez-Callejo, Mallory Freeberg, José Garcia-Pelaez, Lena Guillot-Noel, Tobias B. Haack, Mike Hanna, Holger Hengel, Rita Horvath, Henry Houlden, Adam Jackson, Lennart Johansson, Mridul Johari, Erik-Jan Kamsteeg, Melanie Kellner, Tjitske Kleefstra, Didier Lacombe, Hanns Lochmüller, Estrella López-Martín, Alfons Macaya, Anna Marcé-Grau, Aleš Maver, Heba Morsy, Francesco Muntoni, Francesco Musacchia, Isabelle Nelson, Vincenzo Nigro, Catarina Olimpio, Carla Oliveira, Jaroslava Paulasová Schwabová, Martje G. Pauly, Borut Peterlin, Sophia Peters, Rolph Pfundt, Giulio Piluso, Davide Piscia, Manuel Posada, Selina Reich, Alessandra Renieri, Lukas Ryba, Karolis Šablauskas, Marco Savarese, Ludger Schöls, Leon Schütz, Verena Steinke-Lange, Giovanni Stevanin, Volker Straub, Marc Sturm, Morris A. Swertz, Marco Tartaglia, Iris B. A. W. te Paske, Rachel Thompson, Annalaura

npj | genomic medicine IF 7.3

Explore content · About the journal · Publish with us

nature › npj genomic medicine › articles › article

Article | Open access | Published: 26 October 2024

Comprehensive reanalysis for CNVs in ES data from unsolved rare disease cases results in new diagnoses

German Demidov ✉, Burcu Yaldiz, José Garcia-Pelaez, Elke de Boer, Nika Schuermans, Liedewei Van de Vondel, Ida Paramonov, Lennart F. Johansson, Francesco Musacchia, Elisa Benetti, Gemma Bullich, Karolis Sablauskas, Sergi Beltran, Christian Gilissen, Alexander Hoischen, Stephan Ossowski, Richarda de Voer, Katja Lohmann, Carla Oliveira, Ana Topf, Lisenka E. L. M. Vissers, Solve-RD Consortium & Steven Laurie ✉