Single-cell multi-omics data integration powered by PCA-like autoencoders

Thibaut.peyric@inria.fr

Thibaut PEYRIC[1], Anton CROMBACH[1] and Thomas GUYET[2]

[1] Liris, Inria De Lyon, 56 Bd Niels Bohr, 69100 Villeurbanne, France

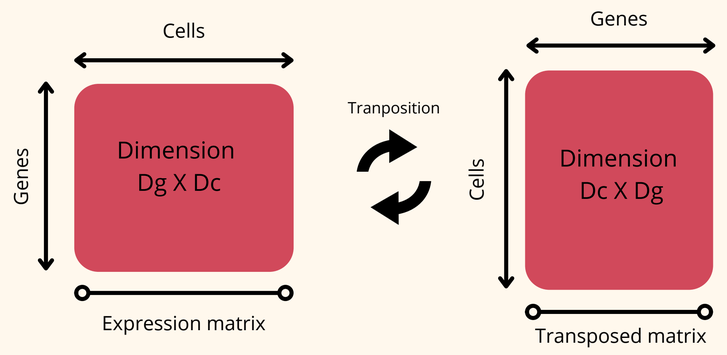
[2] AIstroSight, Inria, Université Claude Bernard Lyon 1, Hospices Civils de Lyon, Villeurbanne, F-69603, France

Introduction

Single-cell (sc) technologies combined with high-throughput sequencing are revolutionizing various omics fields. Each type of omics provides complementary cellular information, but **integrating** these data types remains only partially solved. While integrating data from the same omics or multiple types derived from the same cells is routine, **the current challenge is integrating data from different omics** that do not originate from the same cell population.

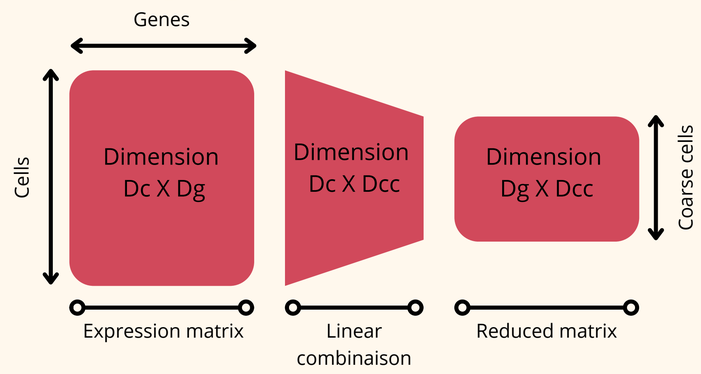
here I present the pipeline based on PCAAE [2] **autoencoder** to learn **gene representation** and integrate scRNA-seq data for multi-omics dataset integration.

1. Matrix transposition



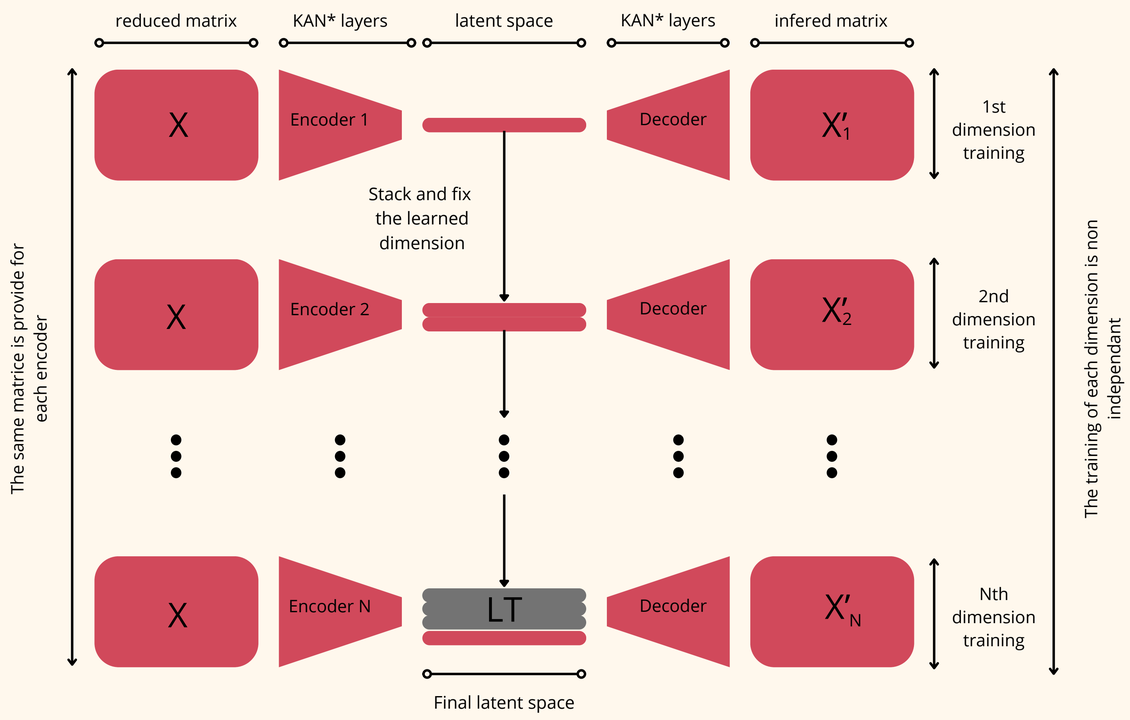
To focus on gene analysis, we transpose the expression matrix so **that the observations are now genes and the features are cells**.

2. Cells agregation to coarse cells



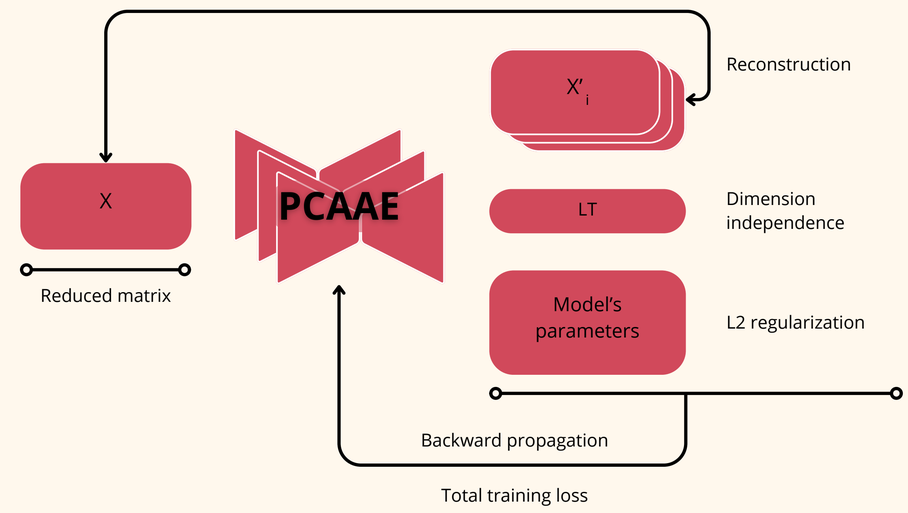
To **limit the number of features encoded** by the model, we perform dimensionality reduction by creating linear combinations of cells.

3. PCAAE auto-encoder architecture



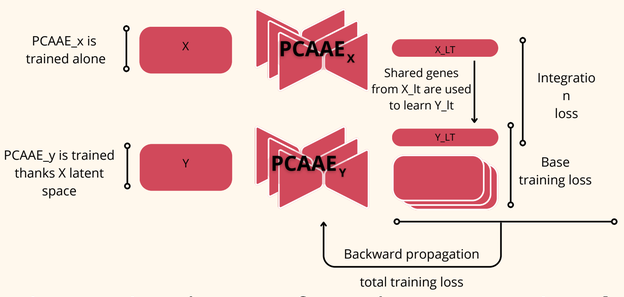
The autoencoder model consists of **N encoders, one for each dimension** in the latent space, and a **single decoder**. Each encoder-decoder pair is **trained sequentially.** Kolmogorov-Arnold Networks (KANs) [3] are use for encoder and decoder as **promising alternatives of Multi-Layer Perceptrons** (MLPs)

4. PCAAE model training



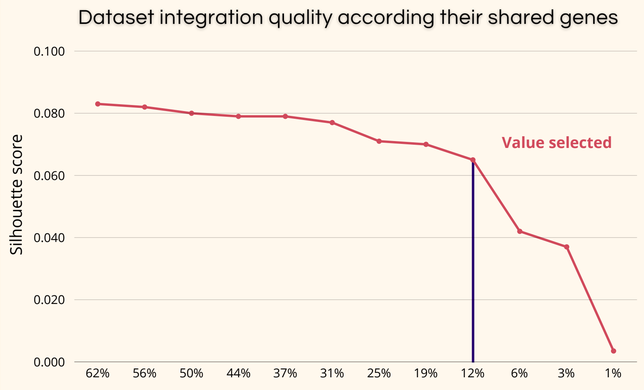
The loss function of the model comprises three terms. The mean squared error (MSE) **facilitates the extraction and compression of data**. The MSE between the covariance matrix of the latent space and the identity matrix **ensures independent dimensions in the latent space**. Finally, L2 regularization is applied to enhance the model's learning process.

5. PCAAE training for integration



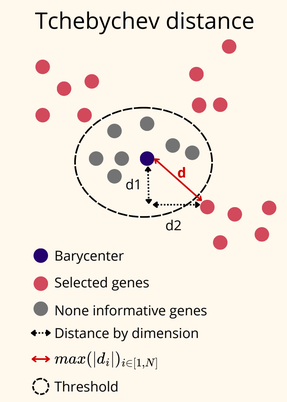
When integrating datasets from the same omics, **the latent** space of one dataset is **used as a reference to train another**. This adds a metric to the model's total loss .

6 . Integration quality



The quality of integration **depends on the percentage of shared genes** used during training. A threshold of 12% was used to demonstrate the quality of the results with the **minimal number** of genes.

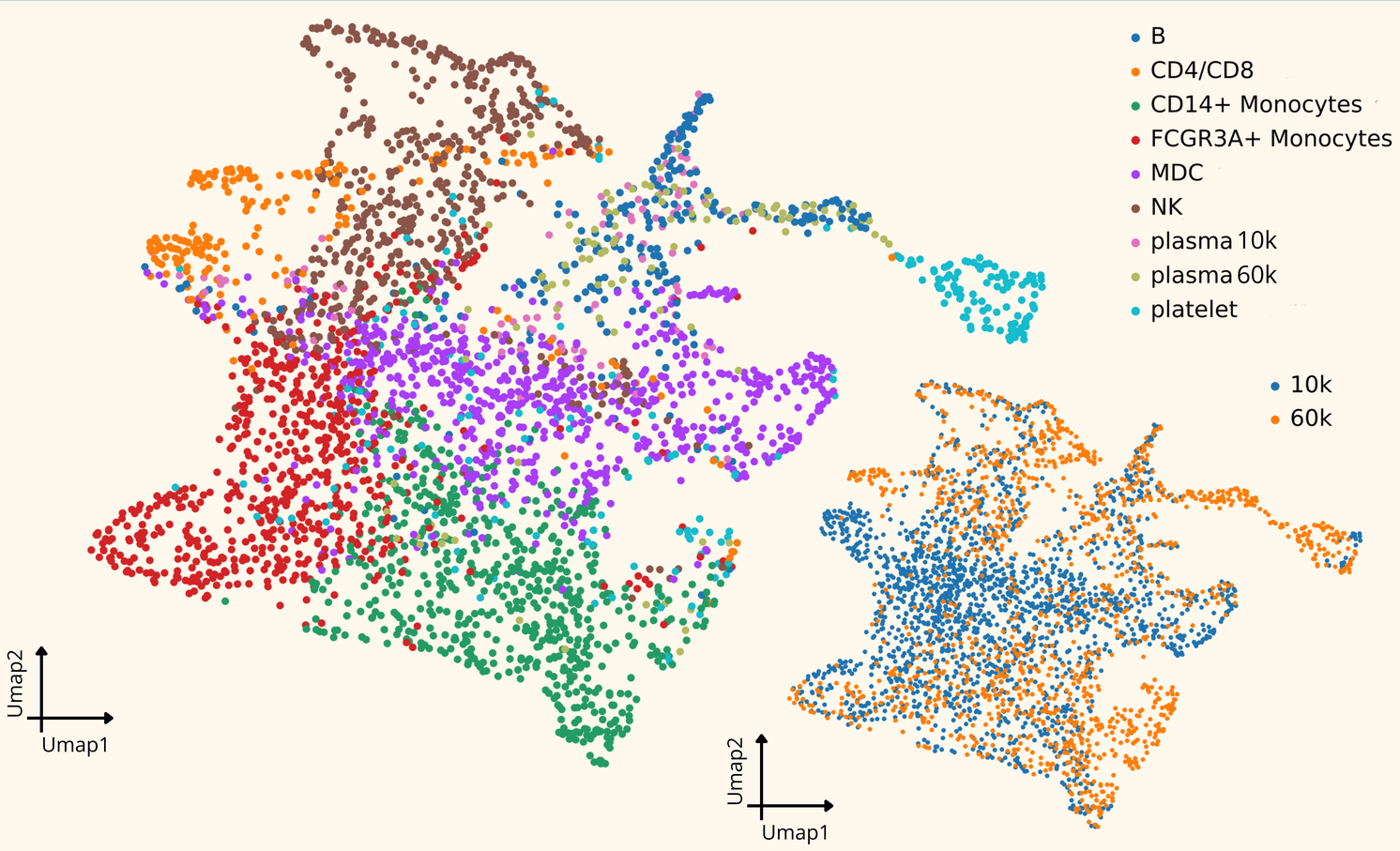
7. Genes selection



After training the model, **gene selection** is performed on the **Tchebychev distance** to the centroid in the latente space. **Genes with high score are considered to contain the most significant information.**

8 . Projection of integrated Genes

UMAP projection of a gene-based latent space. **Each dot is a gene colored by the cell type for which it is most enriched.** Data: PBMC10k and PBMC60K from 10x genomics.



- First results demonstrate **an effective integration** of two scRNAseq datasets, with only 12% of genes used as a link.
- We observe that within the green cluster corresponding to CD14+ monocytes, **genes from both datasets are evenly mixed**
- It indicates that **genes represented are clustered according to their function,** rather than their dataset of origin. Encouraged by this first result
- We are now extending our method to integrate RNA and other omics data (e.g. ATAC).

1. Ghazanfar, S., Guibentif, C., & Marioni, J. C. (2024). Stabilized mosaic single-cell data integration using unshared features. Nat. Biotechnol., 42, 284–292. doi: 10.1038/s41587-023-01766-z
2. C.-H. Pham, S. Ladjal, and A. Newson, PCAAE: Principal Component Analysis Autoencoder for organising the latent space of generative networks, arXiv, (2020).
3. Liu, Ziming, et al. "KAN: Kolmogorov-Arnold Networks." arXiv, 30 Apr. 2024, doi:10.48550/arXiv.2404.19756.

More informations about AI4scMed (project no. 22-PESN-0002)