M2 Research Project

Kevin Dradjat, IBISC, Evry University
kevin.dradjat@univ-evry.fr

2025

Title: Pathway Profiles as Biologically Informed Dimensionality Reduction: A Self-Supervised Learning Approach for Cancer RNA-Seq Data.

Context: High-dimensional gene expression data from cancer samples contain rich molecular information but are difficult to interpret directly. Pathway-based approaches, such as single-sample Gene Set Enrichment Analysis (ssGSEA), transform expression data into pathway activity profiles, summarizing coordinated gene activity into biologically meaningful scores corresponding to cellular processes like proliferation, apoptosis, or immune signaling. These pathway profiles can therefore be considered a form of biologically informed dimensionality reduction, offering a compact and interpretable representation of tumor transcriptomes.

Objective: The core idea of this project is to combine pathway activity prediction and self-supervised learning (SSL) to extract robust and structured representations from bulk RNA-seq data without requiring explicit labels. In SSL, models are trained to solve auxiliary "pretext tasks" that encourage the network to learn meaningful latent features. Common SSL strategies include contrastive learning, where augmented views of the same sample are brought closer in the latent space while different samples are pushed apart, and masked feature reconstruction, where parts of the input (such as subsets of genes) are hidden and the model learns to predict them.

We propose to use pathway profile prediction as a pretext task for self-supervised training. By using this strategy, the model is guided to build a structured latent space that captures biologically relevant representations, without using annotated data.

Methodology: The project will involve preprocessing bulk RNA-seq datasets from public repositories such as TCGA or GTEx, including normalization, gene ID mapping, and feature selection. Pathway activity profiles will be computed using ssGSEA as ground truth for pretext supervision. SSL architectures will be implemented in PyTorch, combining pathway prediction with common self-supervised strategies such as masked-gene reconstruction and contrastive learning. The models will be trained by optimizing the pretext pathway prediction loss. Evaluation will include quantitative metrics such as reconstruction error and correlation with pathway scores, as well as biological assessments using downstream tasks like tumor type or tissue type classification. Visualization of latent spaces using UMAP or t-SNE will be used to interpret the biological structure captured by the embeddings.

References

[1] Kevin Dradjat, Massinissa Hamidi, Pierre Bartet, and Blaise Hanczar. Self-supervised representation learning on gene expression data. *Bioinformatics (Oxford, England)*, 09 2025.

[2] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):9052–9071, December 2024.

[3] Marcelo Segura, Hector Keun, and Timothy Ebbels. Predictive modelling using pathway scores: robustness and significance of pathway collections. *BMC Bioinformatics*, 20, 11 2019.