Biologically Informed Self-Supervised Learning for Gene Expression Data

Svetlana Sannikova
Master 2 GENIOMHE-AI, Univ. Évry Paris-Saclay

université PARIS-SACLAY GRADUATE SCHOOL Informatique et Sciences du Numérique

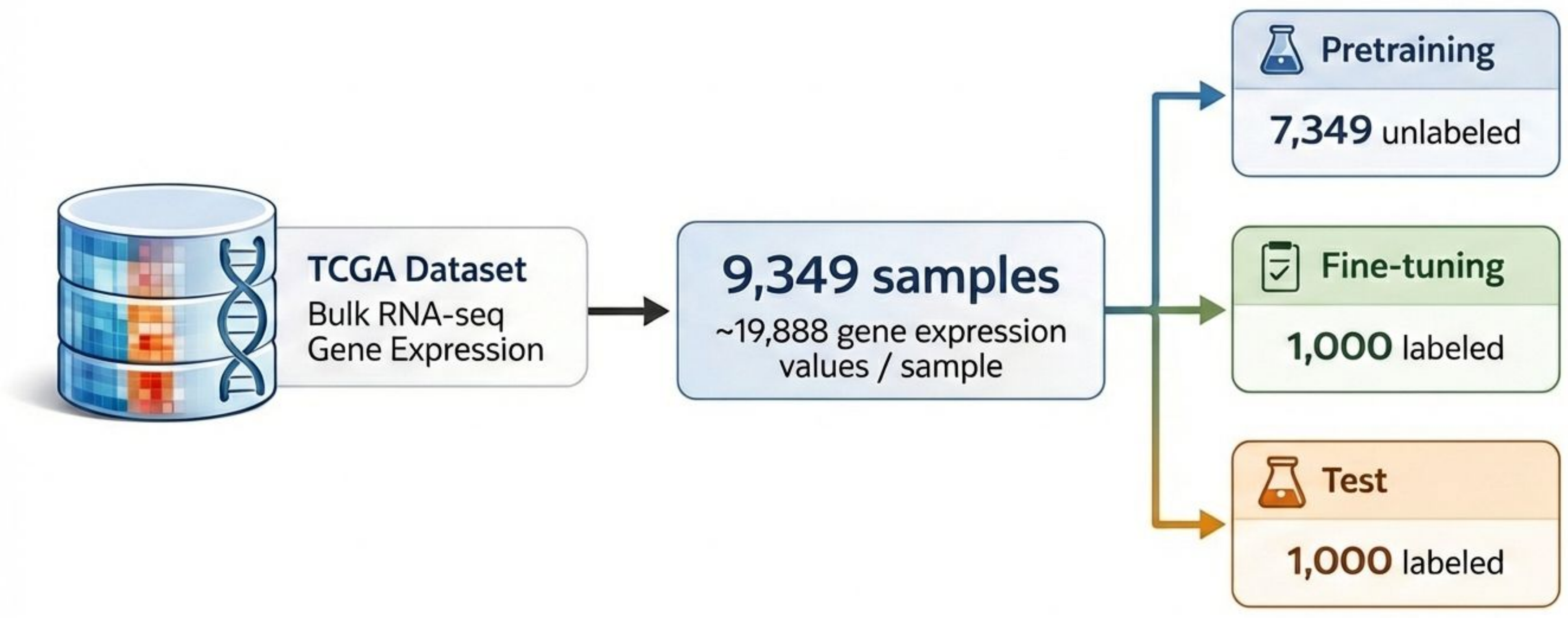
UNIVERSITÉ ÉVRY PARIS-SACLAY

Introduction

Gene expression data are extremely high-dimensional and **weakly labeled**, making supervised learning unreliable in **low-label regimes**.Moreover, standard models ignore the **biological structure of coordinated gene pathways**.

We address this challenge using **biologically informed self-supervised learning** to learn **transferable representations** from **unlabeled transcriptomic data**.

Dataset

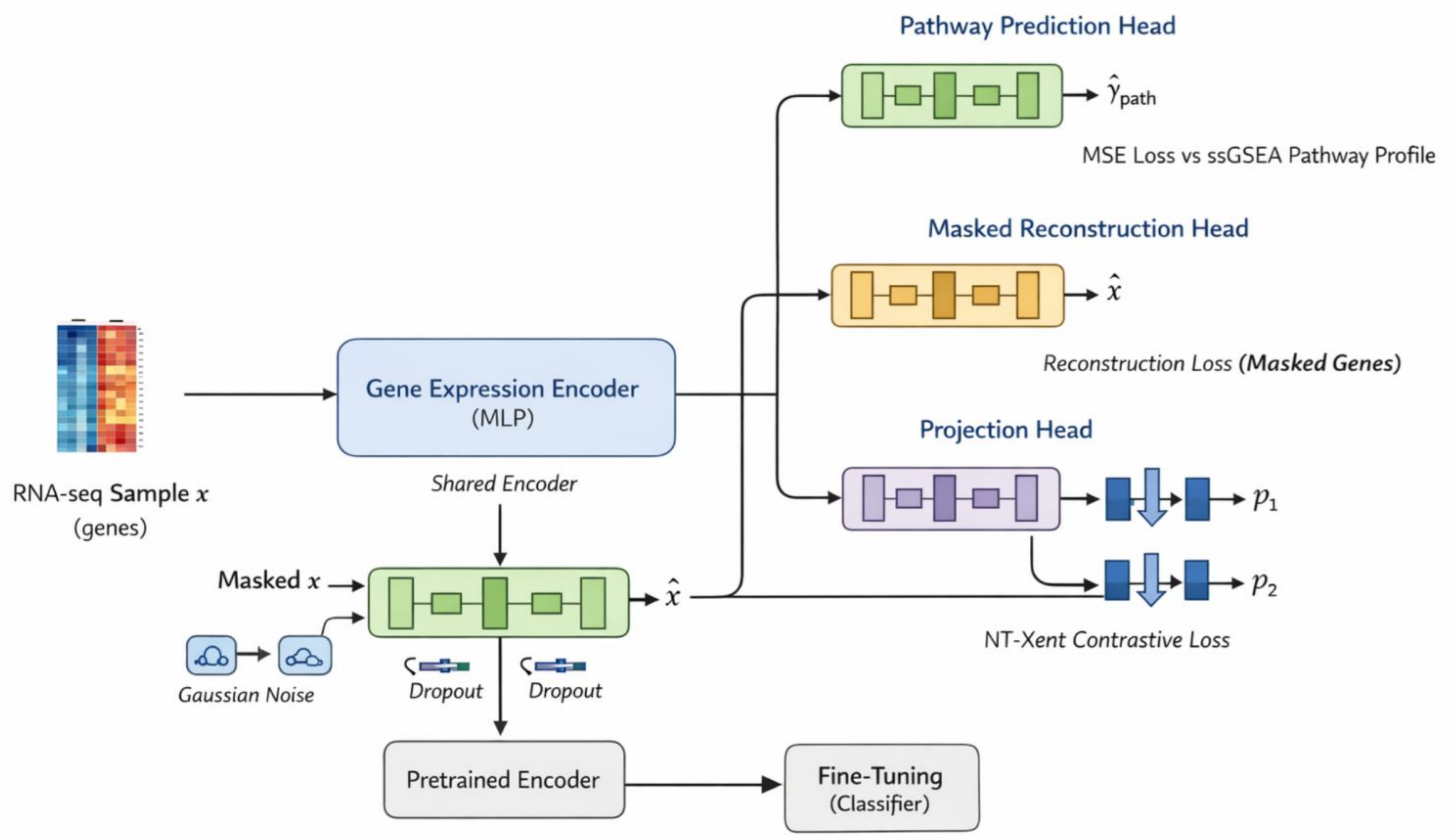


Method Overview

This work presents **self-supervised framework for transcriptomic data** based on a **shared gene expression encoder**.

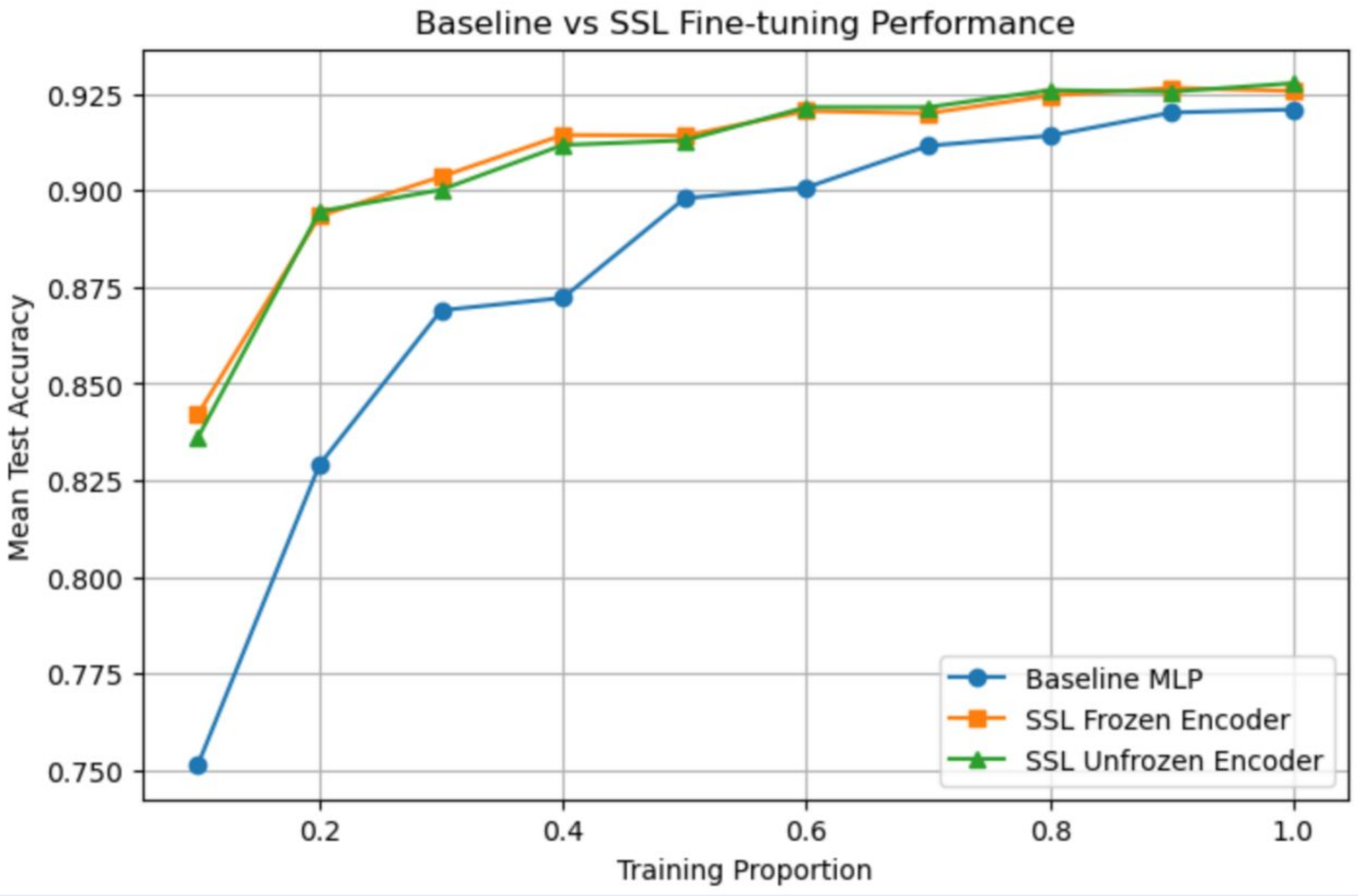
The encoder is pretrained using **pathway activity prediction**, **masked gene reconstruction**, and **contrastive learning**, enabling the model to capture **biologically meaningful and robust representations** from unlabeled RNA-seq data.

The learned representations are evaluated on cancer classification via **linear probing** and **full fine-tuning**, allowing us to assess **data efficiency and transferability in low-label regimes**.



Results

Self-supervised pretraining yields **data-efficient, transferable representation**, with the **largest performance gains in low-label regimes**.



Analysis & Interpretation

Supervised learning degrades sharply in low-label regimes, confirming the difficulty of modeling high-dimensional gene expression with limited annotations.

Pathway-based self-supervised pretraining consistently improves performance, indicating that pathway prediction provides a meaningful biological inductive bias.

Strong results with a frozen encoder show that the learned representations are **transferable** and not task-specific.

Additional gains from full fine-tuning suggest that pretrained representations can be further adapted when more labeled data are available.

Conclusion

Biologically informed self-supervised learning improves data efficiency and transferability in transcriptomic analysis. The learned representations remain effective even with limited labeled data, highlighting their potential for cancer-related applications.

References

[1] **GSEApy developers.** *GSEApy: Gene Set Enrichment Analysis in Python.* https://gseapy.readthedocs.io/

[2] **MyGeneInfo project.** *mygene.py: Python client for MyGene.info API.* https://docs.mygene.info/

[3] **National Cancer Institute.** *The Cancer Genome Atlas (TCGA).* https://portal.gdc.cancer.gov/

[4] **Kanehisa M. et al.** *KEGG: integrating viruses and cellular organisms.* Nucleic Acids Research. https://pubmed.ncbi.nlm.nih.gov/33125081/