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

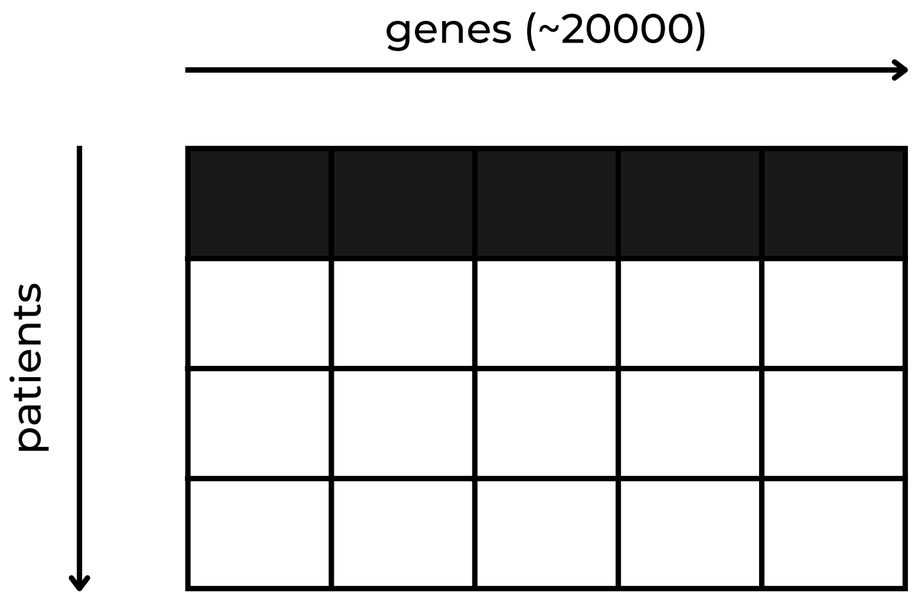
Context :

- Phenotype/cancer prediction from transcriptomic data (bulk-RNAseq)
- Transcriptomic datasets only contains few examples: time-consuming and costly to annotate
- Deep Learning full of potential but application is difficult

Objectives :

- Build a new representation space to enhance the performance of existing models → Foundation Model:
  - Self-supervised Learning (SSL)
  - Pathway profiles
  - Exploit more general unlabeled data

Datasets

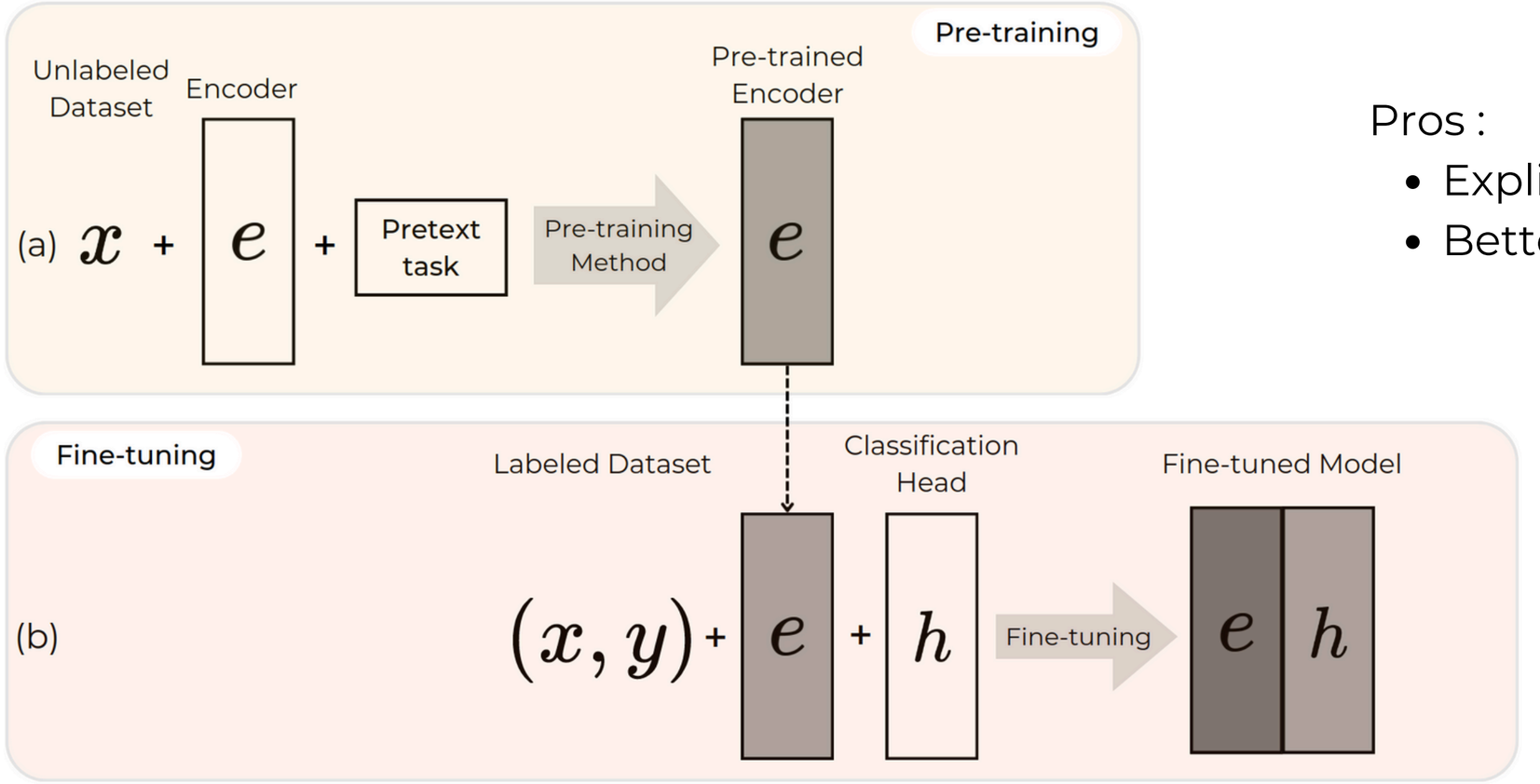


Transcriptomic Data = High-dimensional tabular data

- TCGA (The Cancer Genome Atlas) : Cancer-related dataset
  - 20 types de cancer
  - dimension : (~9500, 20000)
- ARCHS4 (All RNA-seq and ChIP-Seq Sample and Signature Search) : General dataset
  - 10 types de tissus (mostly healthy)
  - dimension : (~45000, 20000)

Self-supervised Learning (SSL)

- Recent advances in computer vision and language (BERT)
- Learn meaningful representations without explicit labels
- Pretext task for the training



Pros :

- Explicit labels not required
- Better generalization

Pathway Profile

Pathway definition: Series of molecular interactions that leads to a specific biological function (ATP production, cellular response, protein production, etc)

Biological pathways:

- From databases: KEGG, MSigDB, Reactome
- one pathway = one gene set (~100-200)
- examples: DNA damage response, apoptosis, immune response, ...

Pathway profile:

- For each sample/patient, we compute an Enrichment Score (ES) for each pathway
- one sample = one vector (representing pathways profile)

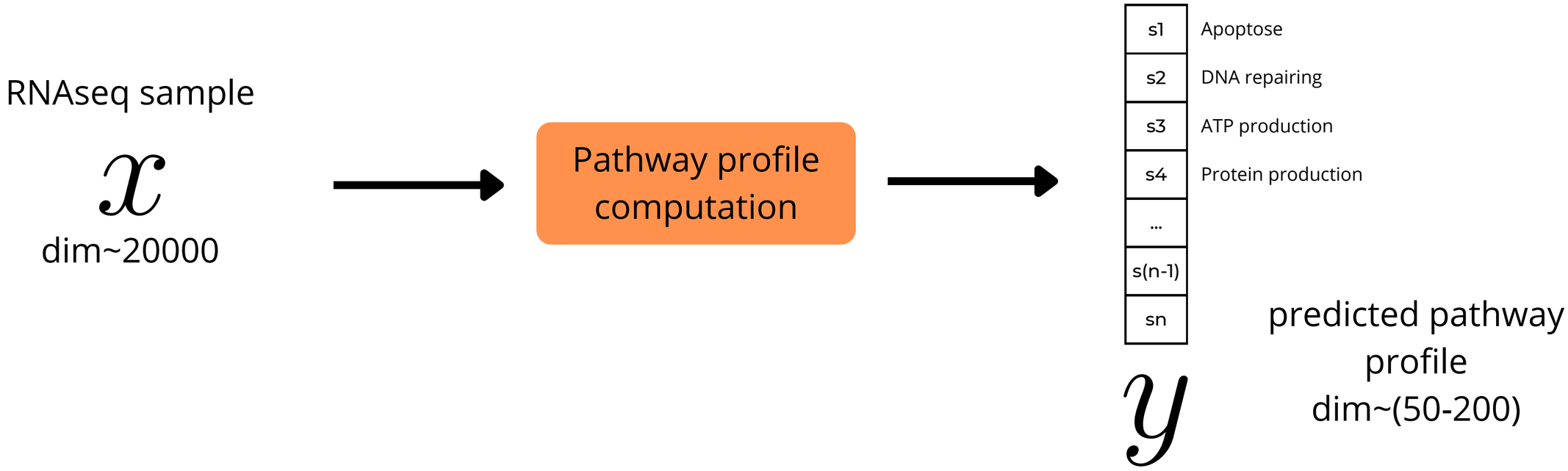
Pathway profile computation pipeline: → Not a simple linear process (GSEA-Gene Set Enrichment Analysis)

- Select a set of pathways
- Rank all genes depending on expression (t-statistic, Log2 Fold, ...)
- Walk down the ranked list to compute the Enrichment Score (ES)

Pathway Profile

Pathway profile:

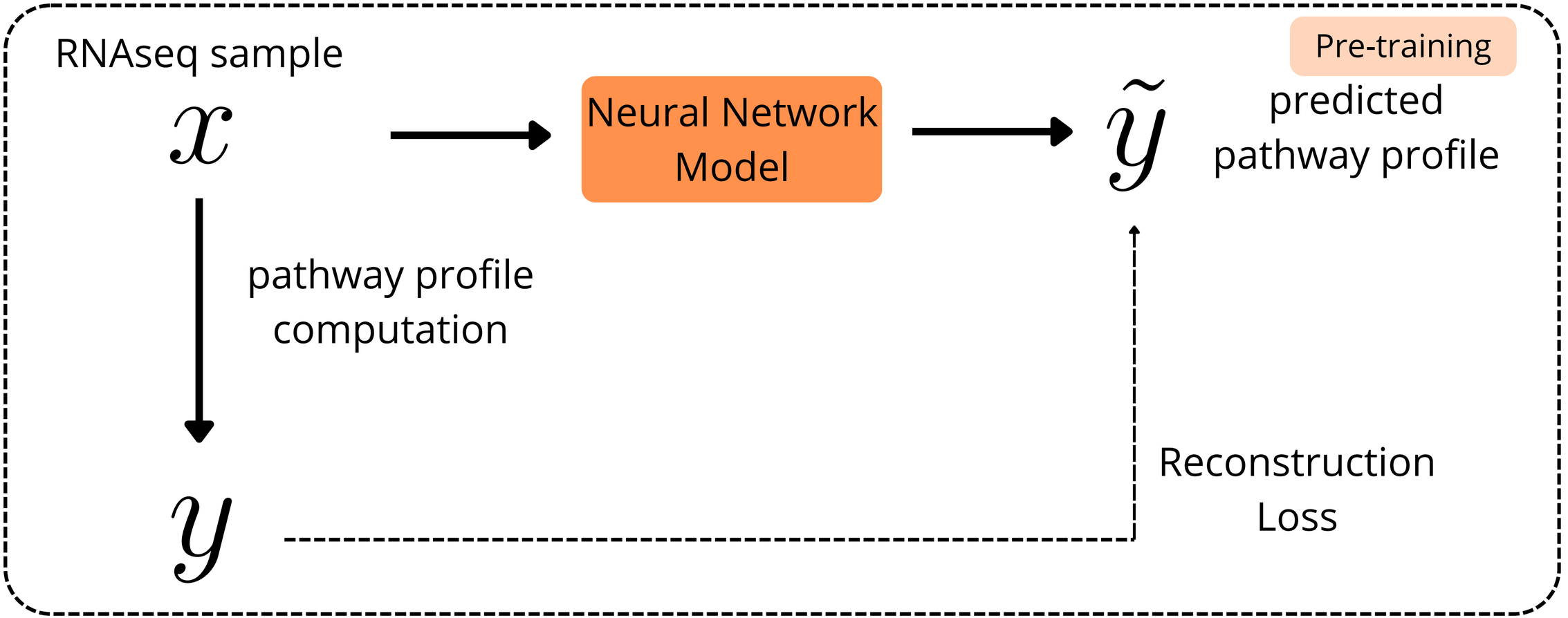
- For each sample/patient, we compute an Enrichment Score (ES) for each pathway
- one sample = one vector (representing pathways profile)



- Can be applied to unlabeled samples
- Can be used as biologically-related dimension reduction / artificial label

Methodology

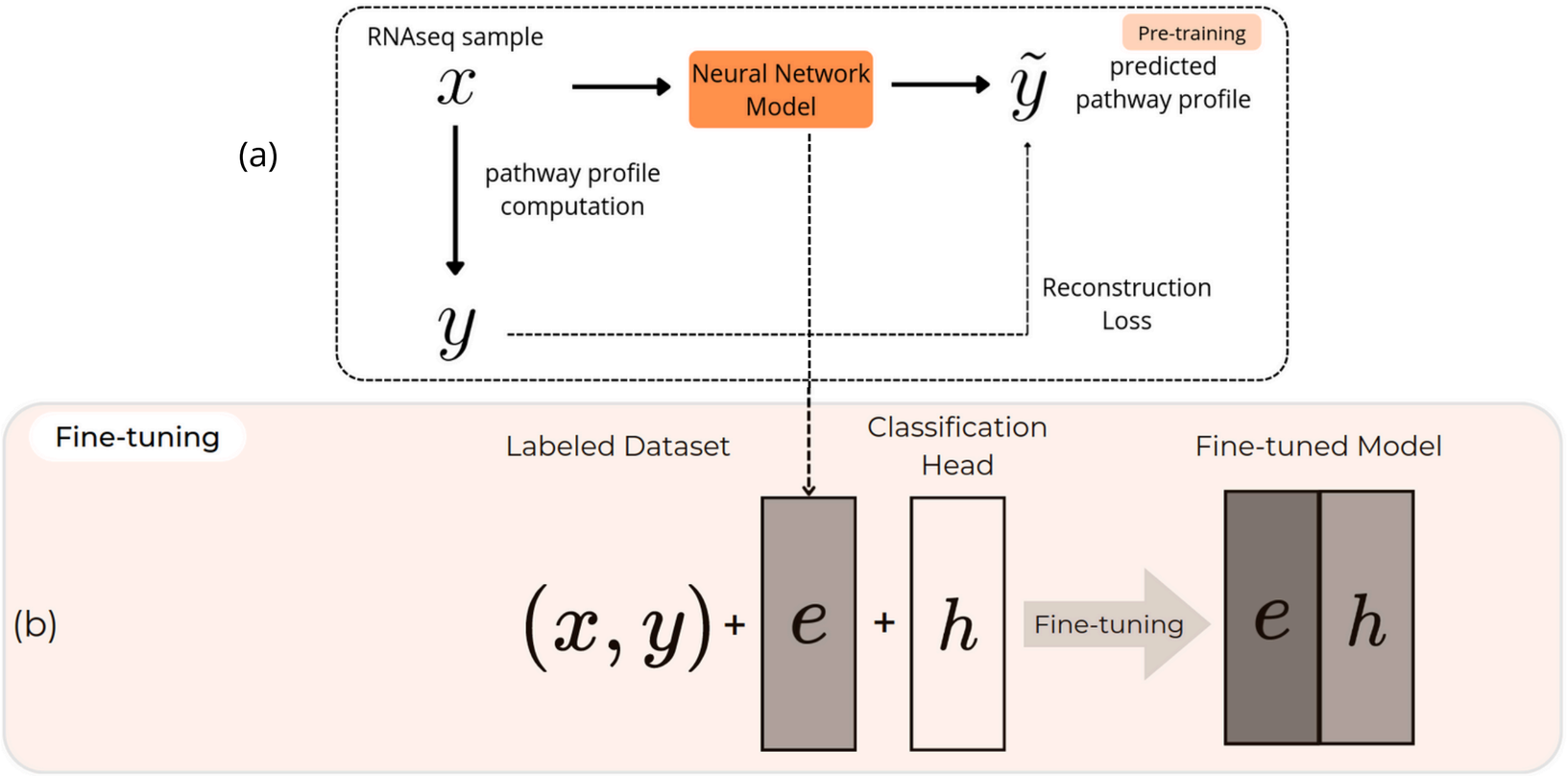
Pretext task: Predict pathway profile from sample



Build a robust foundation model based on biologically-related representation

Methodology

Pretext task: Predict pathway profile from sample



Build a robust foundation model based on biologically-related representation

Methodology

Evaluation:

- Compare pre-trained model to baseline model on a cancer classification task
- Visualize latent space of pre-trained encoder
- Reduce the labeled data used during the fine-tuning