1. Introduction and Problem Statement

Translation efficiency (TE) measures how effectively an mRNA transcript is converted into protein. Ribosomal protein genes (RPL/RPS) are essential for cellular function and exhibit tightly coordinated TE patterns, making them a meaningful target for predictive modeling. Using genome-wide gene expression data to predict TE presents several ML challenges: high dimensionality, correlated features, and potentially nonlinear biological relationships.

This project will build an end-to-end machine learning system that predicts the TE of ribosomal genes using publicly available ribosome-profiling and RNA-seq datasets. The work includes full preprocessing, exploratory analysis, model training, and evaluation. A core objective is to compare classical ML models with more advanced nonlinear and representation-learning methods to identify which techniques best capture structure in TE data.

This project directly aligns with the assignment's example ideas: it uses advanced classical algorithms (e.g., XGBoost, SVR) on tabular data and incorporates neural network models and dimensionality-reduction techniques, consistent with the example encouraging exploration of neural architectures and modern deep learning methods.

2. Data Sources

The project will use publicly available datasets that provide both **RNA-seq expression values** and **ribosome-profiling–based TE measurements**. Suitable sources include:

- Ribosome profiling datasets from the **NCBI Gene Expression Omnibus (GEO)**
- Supplementary data tables from published Ribo-seq studies (e.g., HEK293T)

All preprocessing will be performed within this project, including:

- Importing and aligning RNA-seq and TE data
- Handling missing or low-quality values
- Filtering low-expression genes
- Normalizing and scaling features
- Constructing feature and target matrices
- Creating train/validation/test splits

The specific dataset reference will be included in the final report.

3. High-Level Methods, Techniques, and Technologies

A range of ML approaches will be explored to determine which best predicts TE.

Classical Regression Models

- Linear Regression, Ridge, Lasso, ElasticNet, Principal Component Regression (PCR)

Advanced Classical / Nonlinear Models

- Random Forest Regressor
- Gradient Boosting (XGBoost)
- Support Vector Regression (RBF kernel)

Dimensionality Reduction & Representation Learning

- PCA or Kernel PCA, Autoencoders for latent feature extraction

Neural Network Approaches

- Fully connected feedforward networks. Multi-task regression architectures for predicting multiple TE targets

Tools & Evaluation

- Python, NumPy, pandas, scikit-learn, XGBoost, PyTorch/Keras. Metrics: RMSE, R^2, predicted-vs-actual plots, feature importance, latent-space visualizations

4. Products to Be Delivered

A. Code & Workflow

- Jupyter notebooks for:
    - Data preprocessing, Exploratory analysis, Model training and tuning, Dimensionality reduction, Performance comparison across models

B. Visual Outputs

- Predicted vs. actual TE plots, PCA/latent-space visualizations, Feature-importance charts, Summary performance tables