Final Report: Predicting Translation Efficiency (TE) from RNA Expression Using Machine Learning

1. Introduction and Project Statement

Gene expression is the fundamental process by which genetic information encoded in DNA is ultimately converted into functional proteins. Traditionally, mRNA abundance is used as a proxy for gene activity, largely because it is experimentally accessible and highly informative for many biological processes. However, the abundance of an mRNA transcript alone is insufficient to explain protein output. Protein levels depend not only on transcription but also on how efficiently ribosomes translate each transcript into protein.

This second layer of regulation is captured by **Translation Efficiency (TE)**—defined as the ratio of ribosome occupancy to mRNA abundance. TE measures how effectively ribosomes engage an mRNA and synthesize protein. Factors influencing TE include RNA secondary structure, codon composition, upstream open reading frames (uORFs), translation initiation motifs, ribosome pausing, and global cellular conditions such as nutrient stress or ribosomal biogenesis.

Although RNA levels and TE are correlated, their relationship is imperfect. Some genes exhibit high RNA abundance but low translation efficiency, while others produce abundant protein from relatively few transcripts. This decoupling is biologically important and makes TE prediction both fascinating and challenging.

Biological Motivation

- **High TE genes**: Efficiently translated, often encoding ribosomal proteins or housekeeping factors.
- **Low TE genes**: Tightly regulated; often involved in stress responses, signaling, or development.
- **Ribosomal Protein (RP) genes**: Highly co-regulated group of genes that encode ribosome components; their TE correlates strongly with global translational demand.

Understanding whether TE can be predicted from RNA abundance alone helps clarify:

1. How much translational regulation depends on transcriptional programs
2. How much is governed by sequence-specific, condition-specific, or post-transcriptional mechanisms
3. Whether computational models can reliably infer translational behavior from transcriptomic data

Project Goal

The goal of this project is to evaluate **how well TE can be predicted directly from RNA expression** using a range of machine learning (ML) methods. Specifically, we compare:

- **Linear regression models**
- **Nonlinear ensemble models (Random Forest, XGBoost)**
- **A multi-task fully connected neural network**

By testing models of varying complexity, we determine:

- Whether RNA abundance alone contains predictive information about TE
- Which classes of ML models best capture this information
- What the success or failure of certain models reveals about biological regulation

2. Data Sources and Technologies Used

All data used in this project were generated by **Dr. Can Cenik's Computational Biology Lab** at The University of Texas at Austin, a group specializing in experimental and computational studies of translational regulation. The dataset consists of two matched matrices—RNA expression and Translation Efficiency (TE)—derived from the same set of samples, allowing precise modeling of transcriptional and translational relationships.

Cell Line and Experimental Context

The biological samples come from **HEK293T cells**, a widely used human embryonic kidney–derived cell line characterized by:

- High transfection efficiency
- Robust transcriptional and translational activity
- Stable growth and consistent gene expression patterns
- Frequent use in functional genomics and ribosome profiling studies

Because of their stability and reproducibility, HEK293T cells are ideal for analyzing global patterns of gene expression and translation.

Sample Size and Quality Control

The dataset includes **96 biological samples**, each of which underwent **rigorous multi-step quality-control (QC) filtering** performed by the Cenik Lab. QC processes typically involve:

- RNA integrity checks
- Removal of samples with aberrant sequencing depth

- Filtering out samples with inconsistent ribosome footprint distributions
- Ensuring reproducible mapping rates and gene coverage

Only **QC-pass samples** were included in the final dataset, ensuring that the matrices reflect true biological variation rather than technical noise.

RNA Expression Matrix

- **Dimensions:** 8,433 genes × 96 samples
- Derived from RNA sequencing (RNA-seq) experiments
- Represents steady-state mRNA abundance for each gene
- Includes ribosomal protein genes, housekeeping genes, and diverse functional categories

RNA-seq quantifies the number of transcripts present in each sample. Values were normalized within the Cenik Lab's pipeline to correct for sequencing depth and compositional biases, ensuring comparability across samples.

Translation Efficiency Matrix

The TE matrix contains the **Translation Efficiency values for the same 8,433 genes** in the same 96 samples, derived through ribosome profiling–based quantifications performed in previous Cenik Lab studies.

TE is calculated as:

$$\mathrm{TE} = \log\left(\frac{\text{Ribosome footprint density}}{\text{mRNA abundance}}\right)$$

This ratio captures how actively an mRNA is engaged in translation.

CLR Normalization (Centered Log-Ratio Transformation)

TE values were **CLR-normalized**, a transformation commonly used in compositional data analysis (e.g., sequencing counts). CLR normalization is necessary because:

- RNA-seq and ribosome profiling produce **relative** abundance data
- Counts across genes are interdependent (compositionality problem)
- Uncorrected ratios can distort downstream analyses

The CLR transformation stabilizes variance, reduces spurious correlations, and makes TE values suitable for machine learning algorithms that assume continuous, symmetric distributions.

Matched Gene Ordering for Supervised Learning

A critical feature of this dataset is that:

- The RNA matrix and TE matrix share **identical gene ordering**
- Each row corresponds to the same gene across both datasets
- Each column corresponds to the same HEK293T sample

Because the matrices are perfectly aligned, they can be used directly in a supervised learning framework, where:

- **Input features:** RNA expression values
- **Target outputs:** TE values

This clean alignment removes the need for symbolic mapping, additional preprocessing, or gene reconciliation steps.

Overall, this dataset represents a controlled, high-quality resource for exploring the relationship between transcription and translation.

2.2 Computational Tools and Frameworks

- **Python 3**
- **NumPy, Pandas** for preprocessing
- **scikit-learn** for linear models, Random Forest, PCA
- **XGBoost** for gradient boosting
- **PyTorch** for neural networks
- **Matplotlib** for visualization
- **Jupyter Notebook** for analysis workflow

All modeling steps were implemented modularly with reusable helper functions for loading data, splitting datasets, scaling features, training models, and evaluating performance.

3. Methods Employed

3.1 Preprocessing

Proper preprocessing was essential and heavily influenced model performance. Steps included:

- **Data loading & validation**: Matrices aligned perfectly in dimensions and gene order.
- **Train/validation/test split**: Split across genes (not samples) at 70% / 15% / 15%.
  This tests generalization to *unseen genes*, not unseen conditions.
- **Feature scaling**: Standardized RNA expression using zero mean and unit variance.

- **Quality control**: Included only QC-passed HEK293T samples.

Incorrect scaling previously caused models such as XGBoost and Random Forest to behave unpredictably. After fixing preprocessing, results significantly improved.

3.2 Linear Regression Models

Three linear models were trained:

1. **Ordinary Least Squares (OLS)**
2. **Ridge Regression** (L2 regularization)
3. **Lasso Regression** (L1 regularization; performs feature selection)

These models assume TE is a linear function of RNA expression. Despite their simplicity, they form a strong baseline.

3.3 Nonlinear Ensemble Models

Two nonlinear methods were evaluated:

- **Random Forest Regression**: Ensemble of decision trees; captures local nonlinear interactions.
- **XGBoost Regression**: Gradient boosting machine; powerful for tabular biological data.

These models are capable of modeling complex nonlinear dependencies between RNA expression and TE.

3.4 Dimensionality Reduction: PCA

Principal Component Analysis (PCA) was performed on:

- **RNA expression matrix**
- **TE matrix**

Biological insights from PCA:

- RNA PCA revealed tight clustering of ribosomal protein genes—consistent with strong transcriptional co-regulation.

- TE PCA showed that PC1 captured global translational capacity, while PC2 reflected subtle condition-specific variation.

This supports the idea that TE contains structured, but primarily linear, biological signal.

3.5 Multi-Task Artificial Neural Network

A fully connected neural network (ANN) was built:

- **Input:** 8,433 RNA features
- **Hidden layers:** 512 → 256 (ReLU activations)
- **Dropout:** 0.2
- **Output:** 96 TE values (one per sample)
- **Loss:** Mean Squared Error
- **Optimizer:** Adam

This architecture aimed to capture hierarchical nonlinear relationships not accessible to linear or tree-based models.

4. Results

4.1 Linear Models

Across all evaluated machine learning approaches, the linear models remained the strongest performers. Ordinary Least Squares (OLS), Ridge Regression, and Lasso Regression each explained approximately **53% of the variance in TE**, with RMSE values around 0.80. Specifically, OLS produced an RMSE of 0.8063 with an R^2 of 0.5293, Ridge yielded an RMSE of 0.8072 and R^2 of 0.5297, and Lasso performed slightly better than the other two with an RMSE of 0.8040 and R^2 of 0.5324. The similarities among the models indicate that regularization—whether L1 or L2—has minimal effect on overall performance, suggesting that multicollinearity among features is limited and that the linear structure of RNA expression effectively explains a substantial portion of TE variation.

Biologically, the finding that linear models alone account for roughly **53% of TE variation** is meaningful. It supports the interpretation that TE is strongly influenced by global transcriptional programs, especially for ribosomal protein genes, which tend to be co-regulated and exhibit predictable translational behavior. However, the remaining unexplained variance (~47%) highlights the role of additional post-transcriptional regulatory mechanisms that cannot be captured by RNA abundance alone. These include mRNA secondary structure, codon usage bias, untranslated region (UTR) elements, interactions with RNA-binding proteins, and ribosome pausing dynamics. Thus, while RNA expression is informative, it is far from the complete story of translational control.

4.2 Nonlinear Models

After correcting preprocessing issues such as feature scaling, the nonlinear models performed substantially better than earlier attempts, though they still did not surpass the linear models. The Random Forest regressor achieved a positive R^2 of **0.2734**, showing that it was able to extract some signal; however, its RMSE remained extremely high at **128.26**, reflecting instability and sensitivity to variance. This suggests that Random Forest struggled with the high dimensionality of the dataset (8,433 features) and would likely require extensive hyperparameter tuning—such as limiting tree depth or increasing minimum samples per leaf—to achieve greater stability.

XGBoost, in contrast, produced robust and consistent performance, with an R^2 of **0.4990** and an RMSE of **7.37** after proper scaling. Its R^2 approaches that of the linear models, indicating that although nonlinear relationships between RNA expression and TE exist, they are not dominant. XGBoost was able to model some of these interactions, but such nonlinear contributions appear to be relatively minor compared to the strong linear component already captured by OLS, Ridge, and Lasso.

In summary, while nonlinear models were capable of learning additional structure, their benefits were limited. The results suggest that TE behaves as a mostly linear function of RNA expression across the gene set analyzed, with nonlinear effects contributing only marginally to prediction accuracy.

4.3 Neural Network Performance

Despite correcting preprocessing and scaling issues, the neural network model continued to underperform and did not converge to meaningful predictive accuracy. During training, the RMSE steadily decreased from approximately **481** at epoch 1 to **127** by epoch 96, but the R^2 remained negative throughout the entire training process. This indicates that the model, even while reducing error magnitude, was unable to learn a mapping that generalized well or explained variance relative to the mean predictor.

The poor performance is likely due to several factors inherent to the dataset and model design. The neural network architecture—containing roughly 1.1 million parameters—was far too large relative to the training set of 5,903 genes, resulting in severe overparameterization. Additionally, attempting to predict TE across 96 samples simultaneously introduces a multi-task learning challenge that is difficult to solve with limited data. Deep learning approaches also benefit from sequence-level features—such as codon usage, UTR motifs, and RNA secondary structure—that were not included in this dataset. Without biologically rich features, the neural network had little advantage over simpler models and lacked sufficient information to learn meaningful nonlinear representations.

4.4 Final Model Rankings

When ranked by **R^2**, the best-performing models were:

1. Lasso Regression ($R^2 = 0.5324$)
2. Ridge Regression ($R^2 = 0.5297$)
3. OLS ($R^2 = 0.5293$)
4. XGBoost ($R^2 = 0.4990$)
5. Random Forest ($R^2 = 0.2734$)
6. Neural Network ($R^2 \approx -0.02$)

When ranked by **RMSE**, the ordering was similar:

- Lasso: 0.8040
- OLS: 0.8063
- Ridge: 0.8072
- XGBoost: 7.37
- Neural Network: 127.41
- Random Forest: 128.26

These rankings reinforce the central finding that linear models outperform more complex architectures on this dataset.

4.5 Key Findings

The results clearly demonstrate that **linear models remain the strongest and most reliable predictors of TE** in this RNA expression dataset. They are stable, interpretable, and powerful enough to capture over half of the variance in TE. XGBoost, a state-of-the-art nonlinear method, performed surprisingly well but still did not surpass linear regression. Random Forest models achieved moderate success but suffered from poor RMSE due to instability in high-dimensional feature spaces. The neural network failed to learn meaningful predictive structure, primarily due to insufficient sample size and lack of sequence-level biological features.

Overall, this suggests that TE is governed predominantly by linear transcriptional relationships in HEK293T cells, with nonlinear contributions playing a secondary role.

5. Conclusions

The results of this project demonstrate that RNA expression levels account for approximately **53% of the variation in Translation Efficiency** across 8,433 genes. This indicates a strong linear relationship between transcript abundance and TE, particularly for ribosomal protein genes that are tightly co-regulated at both transcriptional and translational levels. Linear models—OLS, Ridge, and Lasso—outperformed both nonlinear ensemble methods and the deep neural network, confirming that simple statistical approaches remain robust and effective for high-dimensional biological data.

Nonlinear models such as Random Forest and XGBoost were able to capture additional variation, but not enough to outperform the linear benchmarks. This suggests that while TE is influenced by nonlinear factors, these influences are overshadowed by dominant linear regulatory patterns. The neural network, despite its theoretical capacity for learning complex patterns, was unsuccessful due to the limited sample size relative to model complexity, as well as the lack of sequence-based features that are essential for understanding translation.

Scientifically, these findings emphasize that while mRNA abundance is a major determinant of TE, significant additional variation arises from post-transcriptional processes such as codon usage, UTR-mediated regulation, RNA folding, and ribosome pausing—none of which were included in the RNA-only feature set. Future studies incorporating nucleotide sequence features, structural predictions, or ribosome profiling data may enable models, especially deep learning architectures, to surpass the limits observed here.

From a technical standpoint, this work highlights the importance of rigorous preprocessing, proper scaling, and cautious model selection for high-dimensional biological datasets. More complex models do not inherently deliver better results; instead, the model must match the structure and availability of data. Linear regression, despite its simplicity, was the best match for the biological and statistical properties of this dataset.

Future research directions include integrating sequence-derived features, experimenting with Transformer-based sequence models, and developing multimodal architectures that combine expression data with structural or sequence-level inputs. Such approaches may capture the nonlinear regulatory landscape that RNA expression alone cannot fully explain.

6. References

1. Cenik, C. Lab Dataset: RNA and TE measurements from QC-pass HEK293T samples.

2. Ingolia, N. (2016). Ribosome profiling: new views of translation. Nature Reviews Genetics.
3. Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. KDD.
4. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.