HUMBOLDT-UNIVERSITÄT ZU BERLIN

Lebenswissenschaftliche Fakultät

M.Sc. Biophysics

Systems Biology: Computational Analysis and Interpretation of High-throughput Data

Exploring sequence features for TE prediction in different eIF depletion conditions

Zinecker, Anuschka Greta (621202)

Zimper, Oliver (615053)

Förster, Jannes (615480)

March 31, 2026

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CDS coding sequence. 1–3, 11

eIFs eukaryotic initiation factors. 1

EN ElasticNet. 1, 7, 10

LGBM Light Gradient Boosting Machine. 1, 6–8, 10

min ΔG minimum Gibbs free energy. 3, 11

TE translation efficiency. ii, 1–4, 8–11

UTR untranslated region. 1–3, 10, 11

1 Background

The protein abundance of a cell is defined by the three following, interconnected processes: The amount of mRNA transcribed, how efficiently each mRNA molecule is translated into protein, and the rate at which protein is degraded (Zheng et al. 2025). A major regulatory checkpoint of translation in eukaryotic cells is the rate-limiting initiation step, in which the ribosome, assisted by a number of eukaryotic initiation factors (eIFs), is assembled on the mRNA molecule (Marintchev and Wagner 2004). Translation initiation is strongly influenced by cis-acting sequence elements, which are primarily situated upstream of the initiation codon in the 5' untranslated region (UTR) (Tidu and Martin 2024). In order to assess the translation rate, the experimental techniques Ribo-Seq and RNA-Seq can be employed. By normalising the ribosome occupancy, meaning the amount of ribosomes actively translating, to the transcript's steady-state RNA level, the metric translation efficiency (TE) may be used to monitor translation (Ingolia et al. 2009). One approach to disentangle the many regulatory mechanisms determining the translation rate is the controlled depletion of certain eIFs. In this project, two depletion conditions were assessed. The first condition was the knock down of the eIF3d subunit of a larger multiprotein complex. This subunit is crucial for the recruitment and assembly of the translation machinery (Lee et al. 2016). In the second condition, eIF4e, a component of the translation initiation complex 4F, responsible for recruiting the ribosome to the 5' cap structure of mRNA (Svitkin et al. 2005), was knocked down. This report aims to identify sequence determinants of translation that interact with specific initiation factors. This will be accomplished by training and comparing the results of three machine-learning models: Ridge and ElasticNet (EN) linear models, as well as the tree based Light Gradient Boosting Machine (LGBM) model. Their output predicts translation efficiency ratios (log fold change) across two eIF depletion conditions from different subsets of informative sequence features.

2 Methods

2.1 Data preprocessing

The experimental data provided for this project consisted of 11131 protein-coding mRNA transcripts, for which the sequence, length of total sequence, length of 5' UTR, CDS and 3' UTR as well as the TE score for the two depletion conditions and their corresponding control conditions

was supplied. For a clean handling of data, 19 transcripts with missing values in at least one column, as well as transcripts with illegitimate nucleotides were dropped.

2.2 Defining feature sets

In order to predict TE differences from mRNA sequence, the two ML models were trained on strategically chosen feature subsets. As sequence-encoded features are primary determinants of translation efficiency, they can be used as effective predictors to identify motifs and patterns driving translation. A schematic of the pipeline used to create and evaluate these predictions can be found in Figure 1.

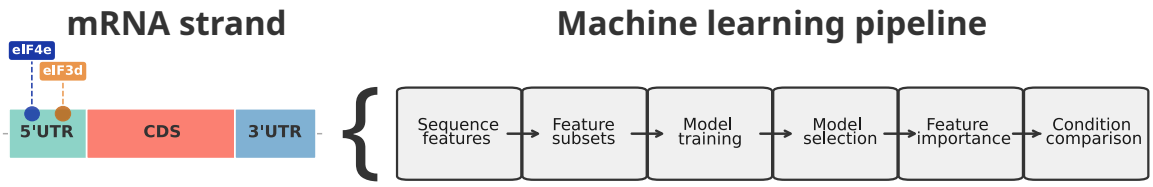

Figure 1: Machine learning pipeline for predicting translation efficiency from mRNA sequence features. (Left) Schematic of an mRNA strand depicting the three major regions — 5' UTR, coding sequence (CDS), and 3' UTR. Translation initiation factors eIF4e and eIF3d are shown bound at the 5' UTR. (Right) Overview of the machine learning pipeline, in which features extracted from the raw mRNA sequence are organized into feature subsets and used to train linear regression (LR) and gradient boosting (LGBM) models. The best performing model is then chosen, features ranked by importance, and results compared across depletion conditions.

The following sequence-encoded features were calculated: The length feature includes the $\log_{10}$ of the 5' UTR, CDS, 3' UTR and total transcript length of each mRNA transcript. Normalised nucleotide frequencies were calculated for the 5' UTR, CDS, 3' UTR and total transcript respectively. Codon and amino acid frequencies were normalized within the CDS. K-mer frequencies were calculated for each of the three regions for k-mers of size 2 through a size of 4. Furthermore, the normalised nucleotide frequency was calculated in the wobble position of each codon (third nucleotide of the codon). All adjacent pairs of codons in each CDS were compared to a set of dicodons found to predict TE in yeast by Gamble et al. 2016. As individual dicodon occurrences are extremely sparse (median count of 2 per transcript, we aggregate all matches into a single count feature normalized by transcript length, to avoid a high-dimensional near-zero feature matrix. To characterize the nucleotide identity surrounding the start codon at the -3, -2, -1, +4

and +5 position, a one-hot encoding for each position was employed. As nucleotide identity at each position is fully determined by the remaining three bases, the T column was omitted from each one-hot encoded position to avoid collinearity. As single strand RNA molecule's collapse into the secondary structure that minimizes their Gibbs free energy ΔG, the SeqFold python package was implemented to calculate the minimum ΔG at physiological temperatures of 37 °C for the sequence of 60 nucleotides starting at the 5' UTR as well as the sequence centered on the start codon to use as predictive features in our models.

Six feature subsets were then defined and used to train the two models and evaluate both, the models' relative performance as well as the contribution of the individual feature subsets to the predictability of the translation efficiency. A full description of features found in each subset is given in the Appendix (Tab.2), we provide a short summary here: The "Basic (no 5'UTR)" subset includes the relative length of CDS, total sequence length, and nucleotide frequency of the CDS and 3' UTR. The subset "Basic" add the corresponding 5' UTR length and nucleotide frequency features. Addition of k-mer frequencies, as well as codon and amino acid frequencies to the "Basic" subset yields the "More Freq." subset. The "Elongation dynamics" subset consists of the nucleotide frequencies in the wobble position of all codons, the nucleotide identity surrounding the start codon and the min ΔG of the transcript. Inhibitory dicodon counts were added to this ensemble to produce the subset "Elongation + Dicodon counts". A "Full" feature subset including all extracted features was also used.

2.3 Linear Models

Both Ridge and Elastic Net linear regression models used in this study predict the log fold-change in TE (control minus depletion) as a continuous response variable $\hat{\mathbf{y}} \in \mathbb{R}^n$ from a feature matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, where $n > 10^4$ is the number of transcripts and p the number of sequence features. Each model is fit independently per depletion condition. The predicted response is modelled as $\hat{y}_i = (\mathbf{X}\boldsymbol{\beta})_i + \beta_0$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the coefficient vector and $\beta_0 \in \mathbb{R}$ a scalar intercept. All features were standardised to zero mean and unit variance prior to fitting so that regularisation penalties act uniformly across features of different scales, with the scaler fit exclusively on the training fold to prevent data leakage. All models were implemented in Python using scikit-learn (Pedregosa et al. 2011).

Ridge Regression

Ridge regression (Hoerl and Kennard 1970) augments the ordinary least-squares objective with an ℓ_2 penalty on the coefficient vector:

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{1}$$

where $\lambda \geq 0$ is the regularisation strength. The ℓ_2 penalty shrinks all coefficients continuously towards zero but does not produce exact zeros, retaining all features in the model. The hyperparameter λ was selected over a logarithmically spaced grid and optimised via 5-fold cross validation.

ElasticNet

ElasticNet (Zou and Hastie 2005) combines the ℓ_1 penalty of the Lasso with the ℓ_2 penalty of Ridge regression:

$$\hat{\boldsymbol{\beta}}_{\text{EN}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \left((1-\alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right) \right\} \tag{2}$$

where $\alpha \in [0, 1]$ controls the mixing between the two penalties and $\lambda \geq 0$ governs the overall regularisation strength. Setting $\alpha = 0$ is equivalent to Ridge regression, while $\alpha = 1$ recovers Lasso. Unlike Ridge, the ℓ_1 component of Equation (2) allows for exact zero-ing of parameters in $\hat{\boldsymbol{\beta}}_{\text{EN}}$, providing implicit feature selection. Both hyperparameters λ and α were jointly optimised by 5-fold cross validation.

2.4 Gradient Boosting Machines (GBM)

Gradient Boosting Machines (GBM) are an ensemble learning method that seeks to build predictive models through the sequential addition of "weak learners" to the ensemble to minimize a loss function $\mathcal{L}(y, F(\boldsymbol{x}))$ (Natekin and Knoll 2013) (He et al. 2019). Given our training data $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^{n}$ with $\boldsymbol{x_i} \in \mathbb{R}^d$ being our feature vectors and $y_i \in \mathbb{R}$ being our "response" variables (in our case the log fold-change in TE), the goal is to obtain an estimate $F(\boldsymbol{x})$ of the optimal function $F^*(\boldsymbol{x})$ mapping $\boldsymbol{x}$ to y that minimizes the expected loss (Friedman 2001):

$$F^* = \arg\min_{F} \mathbb{E}_{\boldsymbol{x}}[\mathbb{E}_y \left(\mathcal{L}(y, F(\boldsymbol{x}))\right) \mid \boldsymbol{x}] \tag{3}$$

In our case the per-sample loss function is the L_2 loss such that the model minimizes the Mean Squared Error (MSE) over the entire training set.

After restricting $F(\boldsymbol{x})$ to be parametric $F(\boldsymbol{x};\theta)$ the boosted model is a weighted linear combination of base learners $h(\mathbf{x};a)$ ("weak learners" producing predictions only slightly better than random guessing) (He et al. 2019):

$$F(\boldsymbol{x};\{\beta_m,a_m\}_1^M)=\sum_{m=1}^{M}\beta_m h(\mathbf{x};a_m) \tag{4}$$

Weights β_m are determined iteratively by updating the model using the step-size ρ_m for each boosting step after an initial guess $\hat{F}_0$ (Friedman 2001):

$$\hat{F}_m(\mathbf{x})=\hat{F}_{m-1}(\mathbf{x})+\rho_m h(\boldsymbol{x},a_m) \tag{5}$$

Using regression trees as base learners the input variable space is divided into regions, identifying those with the strongest responses. Each region is fitted with a regression tree and the mean response of observations is taken. To make the problem of optimizing base learners tractable, functions are chosen to align with negative gradients of the loss function (steepest descent in function space) $\{g_m(\boldsymbol{x_i})\}_{i=1}^N$ along the observed data (Friedman 2001).

$$g_m(\boldsymbol{x_i})=\left[\frac{\partial\mathcal{L}(y_i,F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F=\hat{F}_{m-1}} \tag{6}$$

Parameters a_m are then found through least squares minimization one that fits the base learners after which a line-search is performed to find ρ_m:

$$a_m=\arg\min_a\sum_{i=1}^{N}[-g_m(\boldsymbol{x_i})+\rho h(\boldsymbol{x_i},a)]^2,\qquad \rho_m=\arg\min_\rho\sum_{i=1}^{N}\mathcal{L}\left(y_i,\hat{F}_{m-1}(\mathbf{x}_i)+\rho h(\boldsymbol{x}_i;a_m)\right) \tag{7}$$

$F_m(\boldsymbol{x})$ is subsequently updated as described in Eq. 5.

LightGBM

Conventional gradient boosted decision trees (GBDT) need to scan every data instance for every feature to estimate information gain from split points, making them very time consuming when handling large datasets. LightGBM (Light Gradient Boosting Machine) attempts to address the computational complexity problems of conventional GBDT through the introduction of GOSS

and EFB (Ke et al. 2017).

Gradient-based One-Side Sampling (GOSS). Data instances with larger gradients contribute more to information gain. Through random dropping of instances with small gradients, information content for split finding can be retained even when down-sampling data instances.

Exclusive Feature Bundling (EFB). High-dimensional real world data is usually very sparse, where many features are mutually exclusive (e.g. never taking nonzero values simultaneously). EFB bundles mutually exclusive sparse features to the few dense features, speeding up training without negatively affecting accuracy.

Combined, GOSS and EFB allow LightGBM to scale to large, high-dimensional datasets such as ours, without losing the predictive accuracy provided by conventional GBM. In our implementation TE log fold-change was predicted using the python LightGBM package (Ke et al. 2017) using the LGBMRegressor. Hyperparameters were tuned through a 5-fold cross validation.

3 Results

Despite both linear models achieving similar R^2 on the basic feature set, runtimes differ substantially when scaling to the full 1131-feature set (Appendix Table 3): Ridge's runtime increases 71× to 57 s, while ElasticNet's increases 389× to 38.6 min. LGBMscales more gracefully to 56.3 min (54×). As Ridge regression performed slightly worse than ElasticNet across all feature subsets and depletion conditions, the remainder of this section will examine LGBM and ElasticNet exclusively.

LGBM outperformed ElasticNet regression across all feature subsets and depletion conditions, with average and per run R^2 scores shown in Fig 2. Full model prediction results are found in the Appendix (Table 1), with R^2 values being the averages of the 5-fold CV using a 80/20 training and test data split. Both models achieved the highest R^2 values across both depletion conditions when the full feature subset was included, and performed significantly better on the eIF3d depletion condition when compared to eIF4e, with peak R^2 values of 0.55 and 0.24 respectively. Of note is the significant improvement of predictive power for the eIF3d depletion condition when only the basic (without 5'UTR) feature subset was included ($R^2 = 0.44$) compared with the eIF4e depletion condition ($R^2 = 0.04$) for the LGBM model. Overall LGBM

trained on the full feature subset was the best performing model for both depletion conditions. Scatter plots showing predicted versus observed TE differences for the best performing model

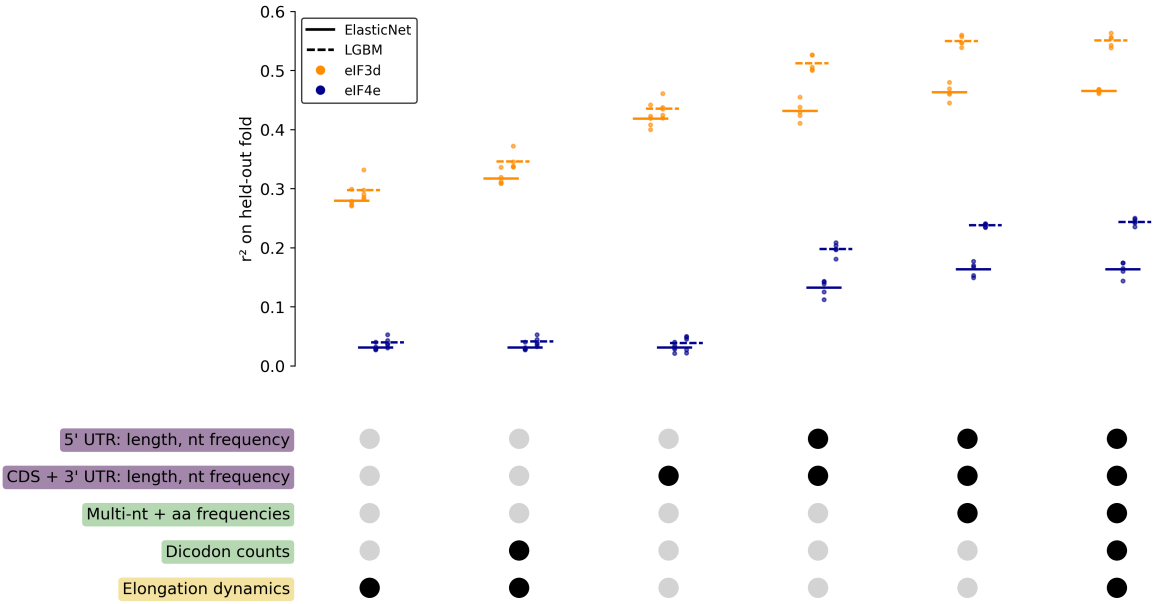

Figure 2: **Model performance across feature subsets and depletion conditions.** Top panel: R^2 on held out folds of ElasticNet and LGBM models trained on different feature subsets, shown for eIF3d (orange) and eIF4e (blue) depletion conditions. Each point represents an individual cross-validation split, horizontal lines indicate the mean performance for each model-condition combination. Feature subsets subsets are ordered from left to right by increasing R^2 (based on EN performance in the eIF3d condition). | Bottom panel: Composition of each feature subset with filled circles indicating inclusion of a given feature group. Feature groups are colour-coded by category: UTR/CDS sequence features (purple), nucleotide/amino acid composition features (green), and elongation-related features (yellow).

(LGBM, full feature set) are shown for both conditions in Fig. 3. The R^2 (averaged across held out folds) was 0.55 and 0.24 for eIF3d and eIF4e, respectively. Pearson correlation coefficients were $r = 0.74$ for eIF3d and $r = 0.49$ for eIF4e. Spearman's rank correlation coefficients were $\rho = 0.76$ and $\rho = 0.52$ for eIF3d and eIF4e respectively. Predicted TE differences aligned better with observed TE differences for the eIF3d depletion condition when compared to the eIF4e condition, as would be expected from the increased performance metrics for both the training and test data.

The ten most informative features were obtained from the importance scores for the LGBM model trained on the full feature set and are shown in Fig 4. Importance values give the total improvement in the model's loss function that a feature contributes when used in a split (Gain based importance). For the eIF3d condition the 5' UTR fraction, meaning the length of the region, was the most important feature, with several other 5' UTR k-mers being among the most

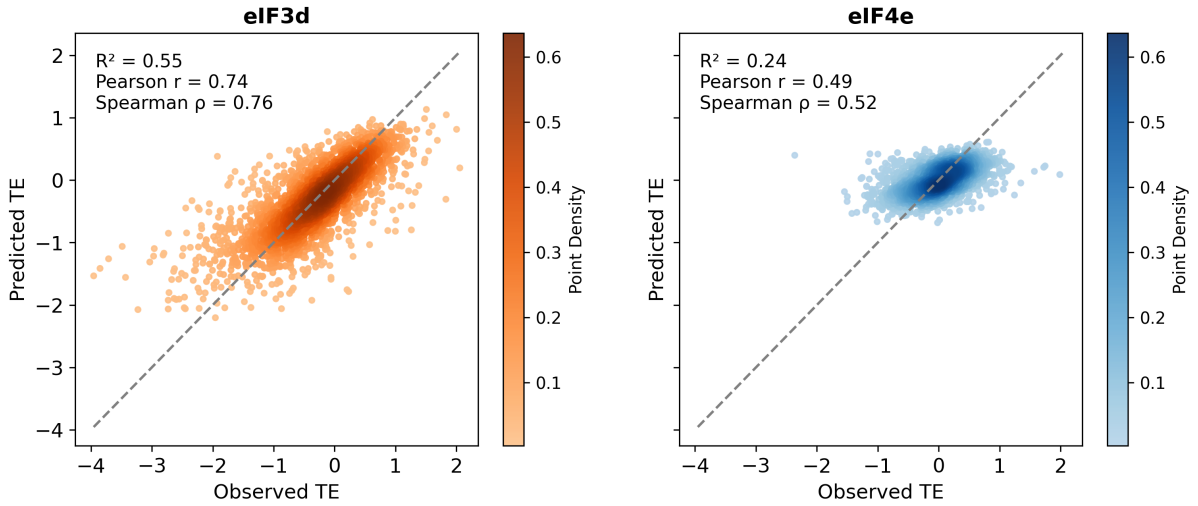

Figure 3: **Predicted versus observed translation efficiency.** Scatter plots of predicted by LGBM trained on subset "Full" versus observed translation efficiency (TE) values for eIF3d depletion (left) and eIF4e depletion (right). The dashed line indicates the identity line (Observed TE = Predicted TE). Reported in each panel are R^2, Pearson correlation r and Spearman rank correlation ρ. n_{eIF3d}=n_{eIF4e}=2223 (one held-out test fold; 20% of 11112 transcripts retained after quality filtering).

informative features. Length features were also highly represented with total transcript length and CDS fraction, or length, also ranking highly. Amino acids frequencies were more commonly represented in the eIF4e condition. 5' UTR and CDS fraction continue to be included, however total transcript length is not included. Additionally elongation dynamics were more included with minimum Gibbs free energy (ΔG) at the start codon and the start of the 5' UTR being highly informative.

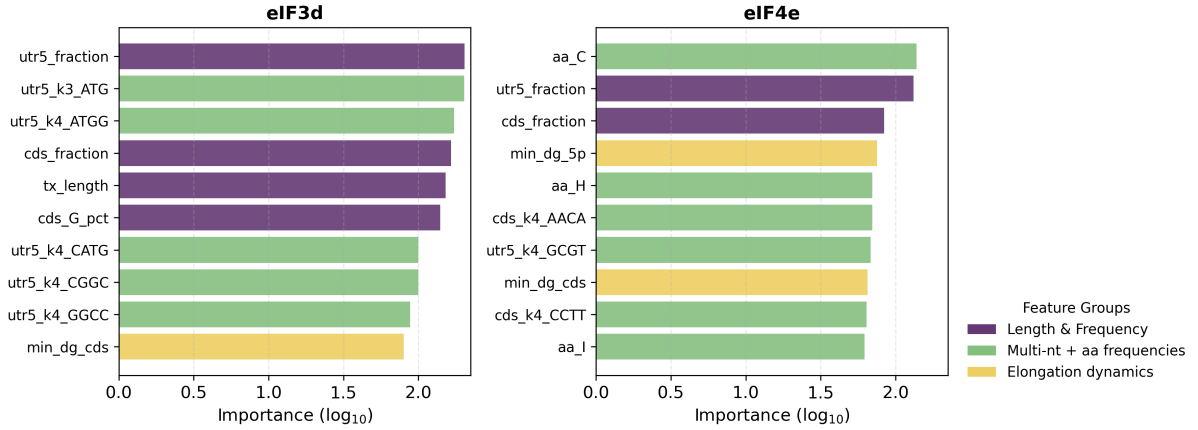

Figure 4: **Top feature importance scores for TE prediction across depletion conditions.** Bar plot showing the top 10 most important features for predicting TE under eIF3d (left) and eIF4e (right) depletion. Feature importance values are $\log_{10}$-transformed. Bars are colour-coded according to feature group: length and nucleotide frequency features (purple), sequence motif and amino acid composition features (green), and elongation-related features (yellow).

A comparison in feature importance is given through a heatmap in Fig. 5. Feature with the largest relative difference in importance scores are highlighted, consisting mainly of k-mers in the 5' UTR.

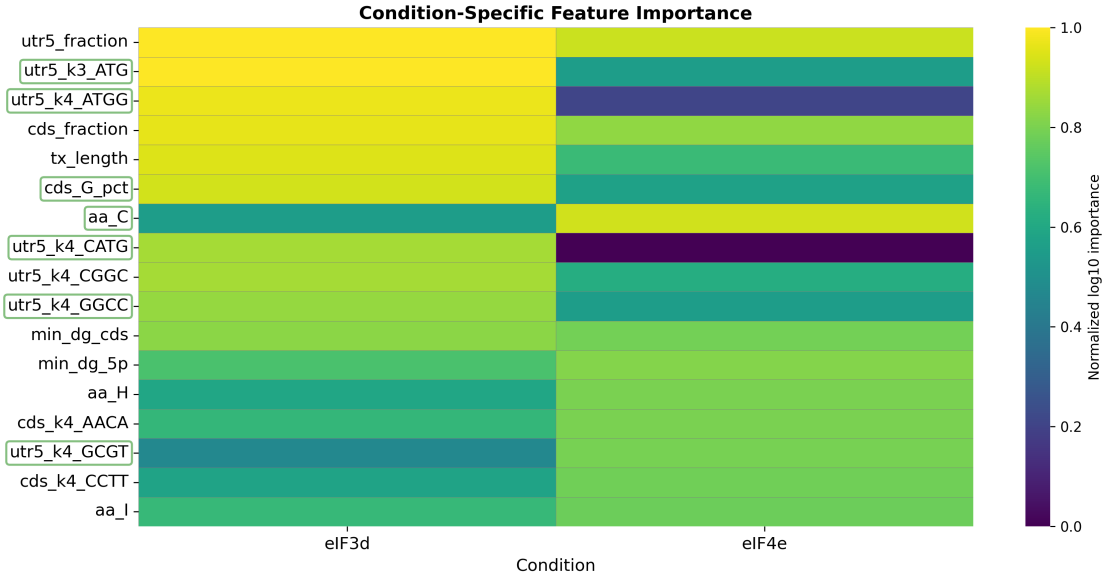

Figure 5: **Cross-condition comparison of feature importance.** Heatmap showing normalised feature importance values across the eIF3d (left) and eIF4e (right) depletion condition. Rows are ordered by mean importance across conditions and features with the largest relative difference between conditions are highlighted according to feature group (green).

The most condition-discriminating features span three feature groups: trinucleotide and tetranucleotide frequencies in the 5' untranslated region (ATG, ATGG, and CATG k-mers, as well as GGCC and GCGT tetranucleotides), guanine nucleotide frequency in the coding sequence, and cysteine amino acid frequency. Overall, the concentration of high-difference features in the 5' UTR is consistent with the broader result that 5' UTR composition is more informative for eIF3d than eIF4e, and points to the 5' UTR as a region of interest for further investigation into condition-specific translational regulation.

4 Discussion

This report aimed to demonstrate traditional machine learning models (ElasticNet, Ridge, LGBM) capabilities in predicting translation efficiency from sequence-derived features and disentangling regulatory mechanisms concerning translation initiation through the depletion of initiation factors. This section discusses the key findings of this report particularly focusing on performance of the models, prediction capabilities between the two depletion conditions, and

implications of feature importance.

LGBM outperformed ElasticNet in TE prediction

Across all feature subsets and depletion conditions, LGBM consistently outperformed ElasticNet in predicting the log-fold change of translation efficiency between control and depletion. This suggests the presence of nonlinear relationships and feature interactions, which linear-based models cannot capture. Zheng et al. 2025 found that LGBM similarly outperformed linear models when predicting TE from sequence-encoded mRNA features across human and mouse cell types, and also demonstrated that the inclusion of amino acid composition information improved LGBM performance more than it did that for linear models. This highlights the importance of nonlinear feature interactions that linear models cannot exploit (Ke et al. 2017). The superiority of tree-based methods over regularized linear regression in biological prediction tasks has been attributed to their ability to model epistatic and nonadditive effects, which regularized linear models fail at (Abdollahi-Arpanahi et al. 2020).

Limited sequence-based predictability of eIF4e compared to eIF3d depletion

Both models achieved a substantially higher predictive power of the TE log-fold change in the eIF3d depletion condition. This may imply, that the eIF3d initiation complex's translation regulation is heavily encoded in sequence-derived features. In contrast, the weak predictability of the TE in the eIF4e condition may point towards regulatory mechanisms less dependent on or not captured in sequence-derived features, like trans-acting mechanisms. Additionally, smaller TE changes in the eIF4e depletion condition against the control condition are reflected in a lower signal-to-noise ratio, which in turn leads to a weaker model performance.

Excluding 5' UTR features strongly reduces predictive performance

The significant drop when excluding the 5' UTR length and frequency features concerning the feature subsets "Elongation Dynamics", "Elongation + Dicodon counts" and "Basic (no 5'UTR)" in both depletion conditions (Fig.2) indicates a critical regulatory role of the 5' UTR region, especially regarding the near-complete predictive loss in the eIF4e depletion (R^2= 0.04). The central role of the 5' UTR in ribosome recruitment and translation efficiency has repeatedly been shown in previous work (Araujo et al. 2012).

Motif-driven features dominate eIF3d, while structural features influence eIF4e predictions

In the eIF3d depletion condition 5' UTR features, like fraction and k-mers, strongly contributed to predictability of the change in TE upon depletion, as well as length features, like total length of transcript and CDS fraction. This suggests that the recruitment and assembly of translation machinery by eIF3d depends heavily on the recognition of sequence motifs. In the eIF4e depletion condition, the feature importance of the min ΔG at the 5' UTR and CDS as well as high scoring of length features, may point towards the eIF4e functionality relying more strongly on structural features of the transcript than sequence-based ones. Similar speculations regarding the role of the secondary structure on the binding preference of eIF4e have been made in studies such as one from Cawley and Warwicker 2012.

5' UTR k-mers prove to be a key discriminative feature

As shown in Fig.5, 5' UTR k-mers have the strongest divergence in feature importance across the two conditions. Notably, three of the highlighted features in Figure 5 (the trinucleotide ATG and the tetranucleotides ATGG and CATG) share an ATG motif. The consistent appearance of these overlapping k-mers across multiple window sizes suggests that the frequency of start codon-like sequences in the 5' UTR is a genuine predictive signal for eIF3d sensitivity, though the overlapping nature of k-mer features means their individual contributions cannot be cleanly separated.

Highlighted in this report is the potential of combining perturbation-based experiments with machine learning models to disentangle regulatory mechanisms controlling the dynamics of translation. It demonstrates that the translation efficiency in regard to different depletion conditions can be partially predicted from sequence-derived features. However, predictability is heavily dependent on the context it is given. While eIF3d-mediated translation appears to be largely encoded in 5' UTR sequence motifs, eIF4e-dependent regulatory mechanisms may rely more on structural or trans-acting mechanisms, as prediction though sequence-based features proved difficult. These findings show the importance of including further levels of translational regulation, such as secondary structure, protein-RNA interactions or other biochemical parameters, to fully capture the IF-dependent control of translation and improve predictive performance.

References

Abdollahi-Arpanahi, Rostam, Daniel Gianola, and Francisco Peñagaricano (2020). "Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes". In: *Genetics Selection Evolution* 52.1, p. 12.

Araujo, Patricia R et al. (2012). "Before it gets started: regulating translation at the 5 UTR". In: *International Journal of Genomics* 2012.1, p. 475731.

Cawley, Andrew and Jim Warwicker (2012). "eIF4E-binding protein regulation of mRNAs with differential 5-UTR secondary structure: a polyelectrostatic model for a component of protein–mRNA interactions". In: *Nucleic acids research* 40.16, pp. 7666–7675.

Friedman, Jerome H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine". In: *Annals of Statistics* 29.5, pp. 1189–1232.

Gamble, Caitlin E et al. (2016). "Adjacent codons act in concert to modulate translation efficiency in yeast". In: *Cell* 166.3, pp. 679–690.

He, Zhiyuan et al. (2019). *Gradient Boosting Machine: A Survey*. arXiv: 1908.06951 [stat.ML]. URL: https://arxiv.org/abs/1908.06951.

Hoerl, Arthur E and Robert W Kennard (1970). "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1, pp. 55–67.

Ingolia, Nicholas T et al. (2009). "Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling". In: *science* 324.5924, pp. 218–223.

Ke, Guolin et al. (2017). "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems* 30.

Lee, Amy SY et al. (2016). "eIF3d is an mRNA cap-binding protein that is required for specialized translation initiation". In: *Nature* 536.7614, pp. 96–99.

Marintchev, Assen and Gerhard Wagner (2004). "Translation initiation: structures, mechanisms and evolution". In: *Quarterly reviews of biophysics* 37.3-4, pp. 197–284.

Natekin, Alexey and Alois Knoll (2013). "Gradient Boosting Machines, A Tutorial". In: *Frontiers in Neurorobotics* 7, p. 21. DOI: 10.3389/fnbot.2013.00021.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12, pp. 2825–2830.

Svitkin, Yuri V et al. (2005). "Eukaryotic Translation Initiation Factor 4E Availability Controls the Switch between Cap-Dependent andInternal Ribosomal Entry Site-Mediated Translation". In: *Molecular and cellular biology* 25.23, pp. 10556–10565.

Tidu, Antonin and Franck Martin (2024). "The interplay between cis-and trans-acting factors drives selective mRNA translation initiation in eukaryotes". In: *Biochimie* 217, pp. 20–30.

Zheng, Dinghai et al. (2025). "Predicting the translation efficiency of messenger RNA in mammalian cells". In: *Nature biotechnology*, pp. 1–14.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2, pp. 301–320.

A Appendix

A.1 Code availability

The full implementation of the data preprocessing pipeline, feature extraction, and model training is available at github.com/jafoers/IF_TE.

A.2 Contributions

Allocation of different work tasks between members of the group are given as follows:

Zinecker, Anuschka Greta:

- Feature extraction (nt-freq, wobble position)
- Plotting of results
- Writing: **Introduction**, **Discussion**, and Methods

Zimper, Oliver:

- Feature extraction (codon-freq, aa-freq, min ΔG)
- Establishing and obtaining results for the LGBM model
- Writing: **Results**, and corresponding parts of Methods section

Förster, Jannes:

- Feature extraction (length, k-mer, one-hot encoding, dicodons)
- Establishing and obtaining results for the Linear models
- Maintenance and creation of **GitHub and Tables**
- Writing: Corresponding parts of Methods section and parts of Results and Discussion sections

A.3 Supplementary tables

Table 1: **Cross-validation performance (5-fold) for all feature sets and models, predicting eIF3d and eIF4e log fold-change.** MSE is computed as $\overline{\text{RMSE}}^2$. Null RMSE: eIF3d = 0.788, eIF4e = 0.433.

		Ridge			**ElasticNet**			**LightGBM**		
Feature Set	n	R^2	MSE	RMSE	R^2	MSE	RMSE	R^2	MSE	RMSE
Target: eIF3d log fold-change										
Basic (no 5'UTR)	10	0.388	0.379	0.616	0.418	0.360	0.600	0.436	0.331	0.575
Basic	16	0.427	0.355	0.596	0.432	0.353	0.594	0.512	0.295	0.543
More Freq.	1008	0.459	0.335	0.579	0.465	0.332	0.576	0.550	0.269	0.519
Elongation	21	0.280	0.446	0.668	0.279	0.446	0.668	0.298	0.424	0.652
Elongation + Dicodon	23	0.317	0.423	0.650	0.317	0.423	0.650	0.346	0.388	0.623
Full	1131	0.458	0.335	0.579	0.465	0.332	0.576	0.551	0.269	0.519
Target: eIF4e log fold-change										
Basic (no 5'UTR)	10	0.016	0.185	0.430	0.031	0.182	0.427	0.038	0.179	0.423
Basic	16	0.121	0.165	0.406	0.132	0.163	0.404	0.198	0.147	0.384
More Freq.	1008	0.149	0.160	0.400	0.167	0.157	0.396	0.238	0.144	0.379
Elongation	21	0.031	0.182	0.427	0.031	0.182	0.427	0.040	0.178	0.422
Elongation + Dicodon	23	0.031	0.182	0.427	0.031	0.182	0.427	0.041	0.178	0.422
Full	1131	0.152	0.159	0.399	0.163	0.157	0.396	0.244	0.142	0.377

Table 2: **Exact feature composition of each subset used in cross-validation.** Some features were excluded to prevent redundancy and collinearity, since they can be inferred from others (e.g., $\text{UTR3}_{\text{frac}} = 1 - \text{UTR5}_{\text{frac}} - \text{CDS}_{\text{frac}}$).

Feature Set	**Contained features**
Basic (no 5'UTR)	log1p of transcript length, CDS fractional length NT frequencies in CDS, NT frequencies in 3' UTR
Basic	All features in Basic (no 5'UTR), plus: 5' UTR length, 5' UTR fractional length, 3' UTR fractional length NT frequencies in 5' UTR.
More Freq.	All features in Basic, plus: ($k = 2, 3, 4$): normalised frequencies of k-mers computed separately for the 5' UTR, CDS, and 3' UTR. Normalised frequency of each of the 64 codons in CDS. Normalised frequency of each of the 20 standard amino acids in CDS.
Elongation	NT frequency at the third (wobble) position of all codons in the CDS. One-hot encoding of Nucleotide identity at positions −3, −2, −1, +4, +5 relative to the start codon. Minimum Gibbs free energy of the 60-NT window beginning at the 5' UTR start; minimum ΔG of the 60-NT window centred on the start codon (both computed at 37 °C via SeqFold).
Elongation + Dicodon	All features in Elongation, plus: Count of yeast TE-associated dicodons Gamble et al. 2016 in CDS. Dicodon count normalised by total length.
Full	Union of all of the above

Table 3: **Training-and runtime by model and feature set.** Wall-clock seconds for 5-fold nested CV including inner hyperparameter search, ran on a Ryzen 5 1600X 6 Cores @ 3.60GHz. Parameter details can be found in the codebase. Scale factor = full / basic time.

Model	**Basic** ($p = 16$) [s]	**Full** ($p = 1131$) [s]	**Scale Factor**
Ridge	0.8	57	71×
ElasticNet	5.9	2314	389×
LightGBM	62.5	3376	54×