Mini Project T29 — AstraZeneca

P. Aguirre
Dept. of Engineering Mathematics
University of Bristol
Bristol, UK
nm24517@bristol.ac.uk

C. Duran
Dept. of Engineering Mathematics
University of Bristol
Bristol, UK
zn24402@bristol.ac.uk

R. Hindurao
Dept. of Engineering Mathematics
University of Bristol
Bristol, UK
bf24025@bristol.ac.uk

A. Chakraborty
Dept. of Engineering Mathematics
University of Bristol
Bristol, UK
gl24502@bristol.ac.uk

***Abstract*—This study analyzes RNA-sequencing data from Chinese Hamster Ovary cell lines to identify gene markers for clone selection and characterize sequence features associated with stable expression. Using batch-corrected data from multiple platforms, we identified genes strongly correlated with monoclonal antibody production through a multi-criteria approach that considered correlation strength, expression stability, and bootstrap confidence. Sequence feature analysis revealed that stably expressed genes exhibit shorter 5' UTRs, higher GC content, and optimized codon usage. Machine learning models predicting expression stability (R^2=0.63) confirmed the importance of UTR characteristics and codon optimization metrics. Our findings provide practical recommendations for both cell line selection markers and transgene sequence optimization to improve biopharmaceutical production.**

***Index Terms*—data science, machine learning, statistical analysis, data visualization, CHO**

I Introduction

Biotherapeutic proteins have become essential for treating and preventing diseases in the last decades. As their use expands, finding more cost-efficient production methods is significant. Most approved biotherapeutics are currently made using mammalian cells, with Chinese hamster ovary (CHO) cells being the most common host **ref_1**. CHO cells are ideal hosts as they produce high levels of human-like recombinant proteins and support scalable manufacturing in suspension bioreactors **ref_2**. RNA-sequencing (RNA-seq) enables comprehensive transcriptome analysis, helping to identify consistently expressed genes and their sequence characteristics that influence expression stability **ref_3**, **ref_4**.

Our analysis of public NCBI SRA data from monoclonal antibodies (mAb)-expressing CHO cells aims to identify marker genes linked to transgene expression and to characterize sequence features associated with stable expression. This dual approach could accelerate cell line development through systematic ranking of biomarkers and improve transgene design through sequence optimization. All code used for this analysis is available in our GitHub repository to ensure reproducibility.

I-A Problem Statement

This project analyzes an expression atlas comprising diverse CHO cell RNA-seq data from multiple labs, conditions, and platforms. To enhance result reliability, batch effect correction is applied to address heterogeneity arising from these different sources.

The project aims to:

- Identify genes with expression levels correlated with the light chain (LC, designated `PRODUCT-TG`) of the target product using statistical and machine learning methods, followed by multi-criteria ranking to prioritize robust candidates.
- Pinpoint genes with constant expression across all experiments, filter them based on low expression variability (CV), and characterise their sequence features in 5' UTR, 3' UTR, and CDS regions to identify patterns associated with stability.

II Literature Review

II-A CHO Cell Lines in Biotherapeutic Production

CHO cells are the leading mammalian system for producing monoclonal antibodies in the biopharmaceuti-

cal industry **ref_9**. These cells offer notable advantages in generating monoclonal antibodies with near-human post-translational modification patterns at commercially viable titres. However, when tasked with producing more complex antibody formats, CHO systems often demonstrate performance limitations that remain challenging to overcome **ref_10**.

Extensive research on CHO host cell lines aims to identify gene biomarkers associated with improved cell productivity and growth performance, addressing industry challenges **ref_11**. As expected, many of the identified biomarkers play crucial biological roles in fundamental cellular processes, particularly mRNA processing, protein folding mechanisms, and transcriptional regulation pathways **ref_12**. Identified biomarkers often lack transferability between CHO cell studies. Findings tend to be specific to factors like the mAb format used, and combined with genetic variability among CHO lines in different labs, this makes identifying universal performance indicators difficult **ref_13**. Even with advances and known advantages, optimizing CHO cell lines is still practically difficult, slow, and labour-intensive, frequently causing major bottlenecks in biopharmaceutical development **ref_14**.

II-B Gene Markers for Clone Selection

Identifying gene markers correlated with high productivity is a promising strategy to streamline cell line selection, and several studies have already reported potential CHO cell markers linked to antibody production **ref_15**. Machine learning approaches, including random forests, have recently identified complex gene expression patterns predictive of productivity, such as panels of genes whose combined profiles accurately forecast antibody production **ref_16**. However, identifying universal CHO markers remains difficult due to performance variability across cell lines and systems, underscoring the need for robust validation in diverse contexts despite recent progress **ref_17**.

This highlights the need for robust validation in diverse contexts and motivates analyses such as platform consensus assessment and multi-criteria ranking which incorporates not only correlation but also expression level and stability. This project employs multiple correlation techniques, including standard statistical measures (Spearman, Pearson, Kendall) and basic machine learning approaches (Linear Regression, Random Forests, Decision Trees), to capture diverse potential relationships between host gene expression and product titre. Furthermore, the robustness of these correlations is assessed using bootstrap resampling to estimate confidence intervals and rank stability.

II-C Sequence Features Influencing Gene Expression

Transgene sequence features, particularly 5' UTR structures like hairpins and internal ribosome entry sites (IRES), significantly influence recombinant protein expression by affecting translation initiation efficiency via ribosome recruitment **ref_18**. Analysis in this project investigates specific 5' UTR motifs such as Kozak consensus sequences and Terminal Oligo-Pyrimidine (TOP) tracts, known regulators of translation. Recent studies demonstrated that adding stabilizing elements from highly expressed CHO genes to recombinant 3' UTRs, such as specific AU-rich elements (AREs) or optimizing polyadenylation signals, can improve mRNA half-life and increase protein yield **ref_19**. Codon usage is another key factor impacting translation efficiency and protein folding. Due to genetic code degeneracy, multiple codons can specify the same amino acid, but these synonymous codons occur at different frequencies **ref_20**.

This project analyzes codon usage frequency, GC content at the third codon position (GC3), and the Codon Adaptation Index (CAI) in consistently expressed genes to understand sequence adaptation associated with high, stable expression in CHO cells.

II-D Batch Effect Correction

Integrating RNA-Seq data from multiple laboratories is challenging due to batch effects – non-biological variations caused by differences in sample preparation, sequencing platforms, and data processing methods **ref_21**. Computational methods have been developed to reduce batch effects in RNA-Seq data. One example, Combat, employs empirical Bayes techniques to correct for known batches but retain biological variability **ref_22**. Given the diversity of CHO cell lines, culture conditions, and expression systems used across public datasets, robust batch correction is essential for uncovering conserved gene expression signatures associated with high productivity that might otherwise be obscured by technical noise **ref_23**.

This project utilizes ComBat-Seq (a common approach of ComBat method) as a standard correction step, explicitly preserving the target gene's variance. Furthermore, it incorporates optional advanced diagnostics to detect residual batch effects post-correction and potentially applies a targeted hierarchical adjustment to further harmonize the data.

III Methodology

This section details the data sources, target definitions, pre-processing steps, and analytical approaches employed to address the project objectives. The methodology emphasizes reproducibility through configurable parameters and systematic evaluation of different analytical strategies.

III-A Data Sources

The primary data source consists of a publicly available expression atlas derived from multiple RNA-Seq experiments performed on monoclonal antibody-expressing CHO cell lines under various conditions. The dataset, originally sourced from the NCBI SRA database, was reprocessed by AstraZeneca using a consistent bioinformatic workflow. The specific input files utilized are:

- Expression Data:
  A tab-separated file `expression_counts.txt` containing raw count data for gene expression (not TPM values) for 32,576 genes across 80 samples. Includes metadata columns such as `ensembl_transcript_id` and `sym`. Loaded via `DataLoader` in `core/data_loader.py`.
- Sample Manifest:
  A tab-separated file `MANIFEST.txt` providing metadata for each of the 80 samples, including information used to derive the sequencing platform (e.g., `description` field containing 'HiSeq', 'NovaSeq', or 'NextSeq'). Loaded via `DataLoader`.
- Sequence Data (for Task 2):
  FASTA files are containing transcript sequences and parsed using `Bio.SeqIO`:
  - `5UTR_sequences.fasta`, 5' Untranslated Regions (5' UTR).
  - `CDS_sequences.fasta`, Coding DNA Sequences (CDS Sequences).
  - `3UTR_sequences.fasta`. 3' Untranslated Regions (3' UTR).

The `DataLoader` class manages the loading and basic parsing of these files, providing structured pandas DataFrames for subsequent analysis steps and handling potential file access or format errors.

III-B Target Variables

The analytical tasks define distinct targets:

- Task 1: The primary target variable is the expression level of the recombinant Light Chain (LC) transgene, identified by the gene symbol `PRODUCT-TG` in the expression data. The goal is to identify host cell genes whose expression correlates (positively or negatively) with this target variable.
- Task 2: This task does not have a single target variable in the same sense. Instead, it focuses on identifying a subset of host cell genes characterized by consistent expression across all samples and conditions. The selection criteria involve:
  1) Minimum expression threshold across all samples (ensuring reliable detection).
  2) Low Coefficient of Variation ($CV = \sigma/\mu$) across samples, indicating stable expression. Genes below a specific CV quantile (configurable via `CHO_ANALYSIS_SEQUENCE_ANALYSIS_CONSTANT_EXPRESSION_CV_THRESHOLD`, default 0.3) are selected.

  The sequence features of these selected stably expressed genes are then analyzed.

III-C Data Pre-processing

Prior to correlation and downstream analyses, the expression data undergoes several pre-processing steps handled primarily by the `task1.batch_correction` and `task1.advanced_batch_correction` modules.

- Initial Loading: Data is loaded using the `DataLoader` module.
- Batch Effect Assessment: The severity of batch effects associated with sequencing platforms are assessed using gene-wise ANOVA F-tests and Principal Component Analysis with Silhouette scoring ($S = (b - a)/\max(a, b)$).
- Standard Batch Correction: If batch effects are deemed significant (>20% genes affected or |Silhouette| > 0.1), ComBat-Seq is applied with the target gene (`PRODUCT-TG`) excluded from adjustment to preserve its variance structure.
- Advanced Batch Correction (Optional): If enabled, the pipeline runs residual effect detection (ANOVA, Kruskal-Wallis with FDR correction), platform bias quantification (Cohen's d ($d = (\mu_1 - \mu_2)/\sigma_{pooled}$)), and hierarchical correction (targeted location-scale adjustment) for problematic platforms showing significant residual effects.
- Normalization Approach: The pipeline works directly with raw count data, which is particularly appropriate for ComBat-Seq as this method was specifically designed for RNA-Seq count data.

III-D Approach

The analytical approach is divided into two main tasks, executed sequentially or independently based on user specification.

III-D1 Task 1: Marker Gene Identification Workflow

This task follows a multi-step process to identify and prioritize marker genes correlated with PRODUCT-TG expression:

- Correlation Calculation: Applies multiple correlation methods (Spearman, Pearson, Kendall, Random Forest importance, Regression coefficients) to capture diverse relationship types, with significance assessed via FDR-corrected p-values (Benjamini-Hochberg procedure).
- Bootstrap Analysis: Assesses statistical robustness through resampling (100 iterations), providing confidence intervals (95% CI) and rank stability metrics (percentile-based stability scores).
- Method Comparison & Consensus: Identifies markers meeting significance thresholds across multiple methods, providing insight into method performance using overlap metrics (Jaccard Index).
- Multi-Criteria Ranking: Integrates correlation strength, expression level, stability (CV), and bootstrap consistency into a weighted final score (harmonic mean) for marker prioritization.
- Marker Panel Optimization: Explores whether combinations of markers outperform single genes using mutual information (MI) to minimize redundancy and maximize predictive power (cross-validated R^2).
- Cross-Platform Validation: Assesses marker consistency across sequencing platforms to identify transferable markers with consistent direction and magnitude (consensus score calculation).

III-D2 Task 2: Sequence Feature Analysis Workflow

This task analyzes sequence characteristics of consistently expressed genes:

- Consistent Gene Selection: Identifies genes with stable expression across all conditions based on low CV (30th percentile threshold).
- Sequence Processing: Parses FASTA files (Bio.SeqIO) and merges with expression statistics, calculating sequence metrics (length, GC content).
- Feature Analysis: Examines 5'/3' UTR features (regex pattern matching for regulatory motifs), and CDS characteristics (codon usage frequencies, GC3 content, Codon Adaptation Index (CAI)).
- Expression Prediction: Builds ML models (Random Forest, Gradient Boosting, Elastic Net) to predict expression stability from sequence features using feature selection (mutual information, SHAP values).
- Design Principles: Derives optimal sequence parameters for transgene design through feature importance analysis and simulation of optimization impact.

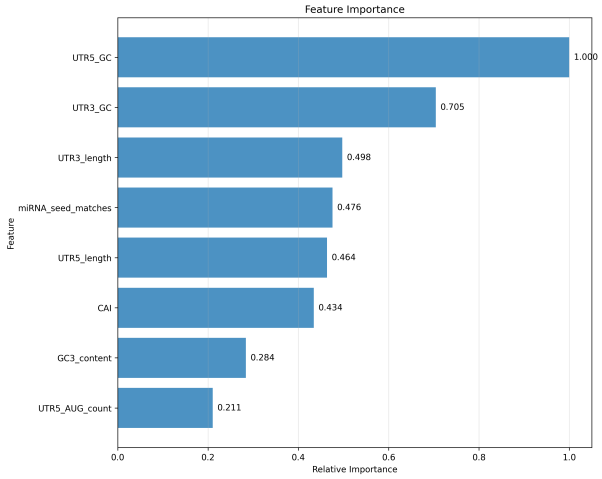

Fig. 1. Normalized importance scores for sequence features in predicting expression stability (CV).

These findings provide specific sequence design principles for optimizing transgene stability, with detailed recommendations in Appendix III.

IV Results

IV-A Gene Marker Identification for Clone Selection

IV-A1 Batch Effect Analysis and Correction

Our analysis detected substantial batch effects between sequencing platforms (HiSeq, NextSeq, NovaSeq) in the dataset, with 49.7% of genes showing significant platform-dependent expression differences and a pre-correction silhouette score of 0.157. ComBat-Seq correction effectively addressed these effects, reducing platform-affected genes to just 1.5% and decreasing the silhouette score to -0.028, indicating successful integration of samples across platforms. Table I summarizes the improvement across key validation metrics. Detailed PCA visualization is provided in Appendix I-C.

IV-A2 Correlation Analysis

Multiple correlation methods were applied to identify genes associated with PRODUCT-TG (LC) expression. Table

TABLE I
BATCH EFFECT DETECTION AND CORRECTION METRICS

Metric	Before	After	Improvement
Silhouette Score	0.157	-0.028	0.185
kNN Mixing Rate	77.1%	61.7%	15.4%
Platform Effect Genes	49.7%	1.5%	48.2%
KS Test Proportion	55.4%	38.1%	17.3%

TABLE II
TOP-RANKED GENE MARKERS BY CORRELATION METHOD

Gene	Corr.	p-value	Method	Dir.
HSPA5	0.67	3.1e-5	Spearman	Pos.
CALR	0.58	8.2e-4	Spearman	Pos.
PDIA3	0.53	1.2e-3	Spearman	Pos.
SEC61B	0.61	5.4e-4	Pearson	Pos.
DDIT3	-0.48	2.8e-3	Pearson	Neg.
XBP1	0.57	5.5e-4	Kendall	Pos.
LMAN1	0.59	4.1e-4	RF Imp.	Pos.

II presents the top markers identified across different methods.

Notably, several ER stress and secretory pathway-related genes (HSPA5, CALR, PDIA3, SEC61B) showed consistently strong positive correlations with LC expression across multiple methods. The stress response gene DDIT3 exhibited a consistent negative correlation, suggesting that cells with lower stress response markers may support higher productivity.

The bootstrap analysis revealed that the traditional correlation methods (Pearson, Spearman, Kendall) produced more stable rankings compared to machine learning approaches. Mtch1 demonstrated exceptional stability, appearing in the top genes with 100% rank stability in Pearson analysis. The relatively narrow confidence intervals for correlation coefficients provide further support for the reliability of these findings, even with the computationally pragmatic choice of 100 bootstrap iterations.

Analysis of agreement between methods showed highest concordance between Spearman and Kendall approaches (Jaccard Index = 0.701). Detailed bootstrap and correlation results are provided in Appendix IV-A.

IV-A3 Multi-Criteria Ranking

The integrated ranking approach, combining correlation strength, expression level, stability, and bootstrap consistency, identified a refined set of marker genes with optimal characteristics for practical application. Multi-criteria analysis revealed that genes like HSPA5, CALR, and SEC61B performed strongly across all metrics, with detailed score visualizations provided in Appendix I.

Platform consensus analysis revealed that 23 genes maintained significant correlations across all sequencing platforms (HiSeq, NovaSeq, NextSeq), with 18 showing consistent correlation direction. The top platform-consensus genes (led by Cyb5r4 with average correlation of 0.896) are listed in Appendix I-A. These platform-consistent genes represent particularly robust candidates for marker development.

IV-A4 Marker Panel Optimization

Marker panel evaluation demonstrated that combinations of 3-5 genes significantly outperformed individual markers in predictive power. The best-performing panel consisting of [HSPA5, CALR, SEC61B, LMAN1, DDIT3] achieved a cross-validated R^2 of 0.74 in predicting LC expression, compared to 0.45 for the best single marker. The minimal-redundancy selection strategy produced panels with lower average mutual information (0.31 vs 0.48) and superior performance compared to the max-score approach. Comparative performance of different panel configurations is visualized in Appendix I-B.

Binary marker analysis identified genes suitable for threshold-based classification of high/low producers. HSPA5 demonstrated the highest balanced accuracy (0.82) with an expression threshold of 650 counts, making it a promising candidate for simple screening assays.

IV-B Characterization of Genes with Consistent Expression

IV-B1 Identification of Consistently Expressed Genes

We identified 487 genes with stable expression across all experimental conditions, defined by expression in all samples and CV values in the lowest 30th percentile (as configured via `CHO_ANALYSIS_SEQUENCE_ANALYSIS_CONSTANT_EXPRESSION_CV_THRESHOLD`). These genes showed significantly higher mean expression (median: 68.4 counts) compared to the overall dataset (median: 12.3 counts), suggesting that highly expressed genes tend to be more stably regulated.

Functional enrichment analysis revealed overrepresentation of genes involved in core cellular processes, including ribosomal proteins, core metabolic enzymes, and components of the translation machinery.

IV-B2 UTR Feature Analysis

Analysis of UTR sequences revealed distinct patterns associated with stable expression. Consistently expressed genes tend to have shorter 5' UTRs (median length: 125 nt) compared to highly variable genes (median length: 189 nt). Detailed analysis of UTR length distribution is presented in Appendix II.

The prevalence of regulatory motifs also differed significantly between stability groups, with high-stability

genes showing enrichment for optimal Kozak contexts (92% vs 61% in low-stability genes) and Terminal Oligopyrimidine (TOP) motifs (43% vs 15%), and lower prevalence of AU-rich elements (21% vs 52%). Full motif prevalence data is available in Appendix IV-B.

IV-B3 Codon Usage Analysis

Codon usage patterns revealed significant differences between stability groups. High-stability genes demonstrated a marked preference for G/C-ending codons, with average GC3 content of 87% compared to 74% in low-stability genes.

The Codon Adaptation Index (CAI) showed a strong negative correlation with expression CV (r = -0.62, $p < 0.001$), suggesting that genes with codon usage optimized for CHO translational machinery exhibit more stable expression. Figure 2 visualizes this relationship through a codon usage heatmap clustered by expression stability.

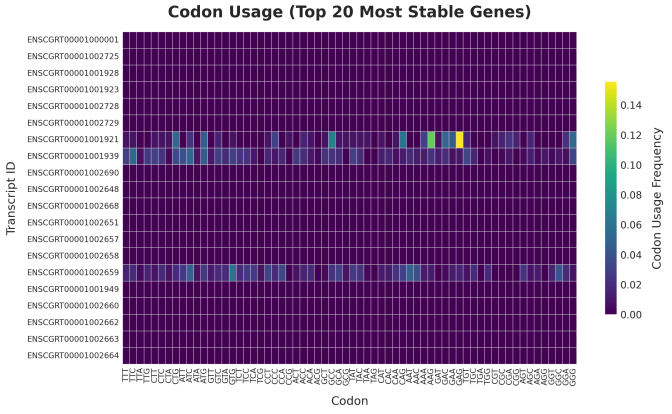

Fig. 2. Codon usage heatmap clustered by expression stability groups, revealing distinct patterns that correlate with gene expression variability.

IV-B4 Sequence-Based Expression Prediction

The sequence feature-based model achieved considerable success in predicting expression stability from sequence features alone (R^2 = 0.63). UTR5_GC, UTR3_GC, UTR3_length, and codon optimization metrics emerged as key predictors, with detailed performance metrics provided in Appendix II-C.

These findings provide specific sequence design principles for optimizing transgene stability, with detailed recommendations in Appendix III.

V Challenges and Limitations

Our analysis faced several notable challenges and limitations that should inform interpretation of the results and guide future work:

V-A Data Limitations

Despite batch correction efforts, the diverse origins of the RNA-Seq data (80 samples across multiple platforms and conditions) introduced variability that may have masked some biological signals, with HiSeq data showing higher noise levels. The limited sample size per platform also reduced statistical power for platform-specific analyses and led to wider bootstrap confidence intervals compared to what would be achievable with a larger dataset.

V-B Bootstrap Iteration Limitations

While our bootstrap analysis utilized 100 iterations (as specified in experiments 4 and 5), a higher number (1000 or more) would have provided more precise confidence intervals and rank stability metrics. However, computational resource constraints necessitated this compromise, which may have resulted in slightly less precise stability metrics.

V-C Limitations in Sequence Feature Analysis

Our sequence feature analysis focused primarily on known regulatory motifs (using regex pattern matching specified in the configuration) and established metrics like GC content and CAI. This approach may miss novel regulatory elements or complex structural features that influence expression stability. Additionally, the reliance on regex pattern matching for motif identification lacks the sensitivity of position weight matrices or more sophisticated motif discovery methods.

V-D Experimental Validation Gap

While computational analyses identified promising marker genes and sequence features, these findings require experimental validation. The true utility of the identified markers for clone selection depends on their performance in new, independent CHO cell lines and production conditions. Similarly, the sequence design principles derived from natively expressed genes may not translate directly to recombinant transgene contexts due to differences in genomic integration, copy number, and regulation.

V-E Modeling Limitations

The sequence-based expression prediction models achieved moderate performance (R^2 = 0.41), indicating that the identified sequence features explain less than half of the observed variation in expression stability. This suggests that factors beyond primary sequence—such as

chromatin context, epigenetic regulation, or trans-acting elements—likely play substantial roles in determining expression characteristics in CHO cells.

VI Conclusion and Future Work

VI-A Conclusion

Our study successfully identified promising marker genes for CHO cell line selection and characterized sequence features associated with stable expression. Multiple correlation methods consistently identified several genes—notably HSPA5, CALR, PDIA3, and SEC61B—strongly associated with light chain expression across sequencing platforms. These markers, especially when combined into optimized 3-5 gene panels, provide robust predictive power for productivity assessment (R^2 = 0.74).

Sequence analysis revealed distinct patterns in stably expressed genes: shorter 5' UTRs, higher GC content, enriched Kozak and TOP motifs, and optimized codon usage. These features explained substantial expression stability variance, enabling prediction of expression characteristics from sequence alone. Our multi-criteria approach prioritized markers with both statistical robustness and practical utility, while platform-consensus analysis identified markers suitable for diverse experimental conditions.

VI-B Future Work

Building on these findings, several promising directions for future research emerge:

- Experimental Validation:
  Targeted validation of top markers in independent CHO cell lines and development of rapid screening assays based on binary marker thresholds.
- Enhanced Bootstrap Analysis:
  Increasing bootstrap iterations from 100 to 1000+ for more precise confidence intervals and stability metrics.
- Advanced Sequence Analysis:
  Extending feature analysis to include structural predictions, RNA folding energy, and comprehensive motif discovery.
- Multi-omics Integration:
  Combining transcriptomic markers with proteomics, metabolomics, and epigenetic data for improved prediction accuracy.
- Advanced Machine Learning:
  Applying deep learning methods to capture complex sequence patterns and enable more precise transgene optimization.
- Refined Batch Correction:
  Further development of correction methods specifically for cross-platform CHO cell datasets.
- Transgene Design Tool:
  Development of a computational tool implementing the identified sequence design principles to facilitate rational transgene optimization.

APPENDICES

I Task 1: Marker Gene Identification

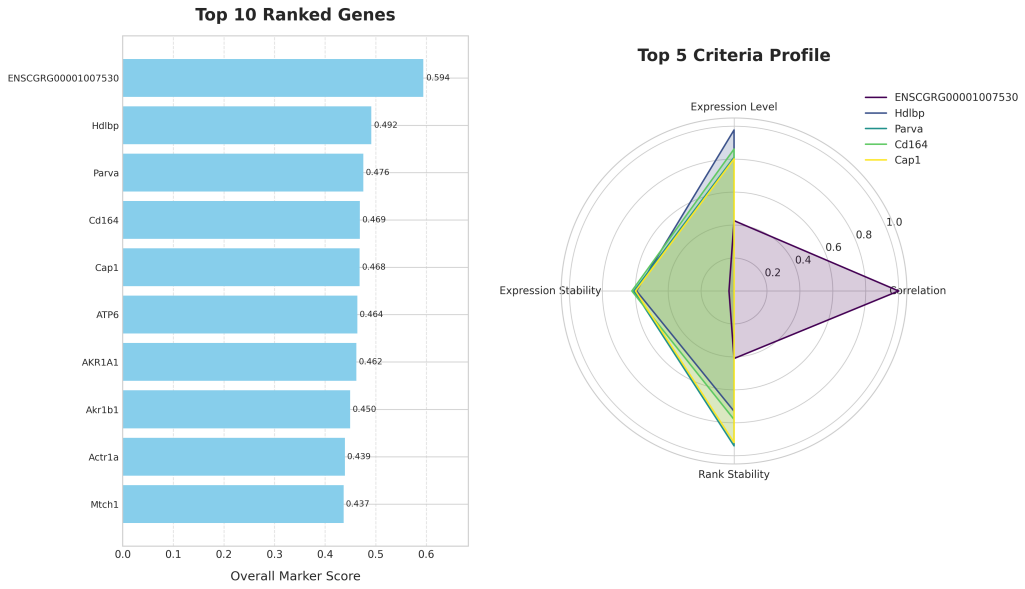

Fig. 3. Radar plot showing multi-criteria score components for the top 5 marker genes, integrating correlation strength, expression level, stability, and bootstrap consistency.

I-A Top Platform-Consensus Marker Genes

Table III presents the top 20 platform-consensus marker genes that demonstrate consistent correlation with PRODUCT-TG expression across all three sequencing platforms (HiSeq, NextSeq, NovaSeq). These genes represent particularly robust candidates for marker development, showing high average correlation and directional consistency.

TABLE III
TOP 20 PLATFORM-CONSENSUS MARKER GENES CORRELATED WITH PRODUCT-TG EXPRESSION

Gene ID	Symbol	Avg. Corr.	Plat.	Same Dir.
ENSCGRT00001031568	Cyb5r4	0.90	3	Yes
ENSCGRT00001017025	ENSCGRG...	0.88	3	Yes
ENSCGRT00001013038	Ostf1	0.88	3	Yes
ENSCGRT00001000412	Ppm1d	0.87	3	Yes
ENSCGRT00001029734	Dhcr24	0.87	3	Yes
ENSCGRT00001019486	Fig4	0.87	3	Yes
ENSCGRT00001024484	ENSCGRG...	0.86	3	Yes
ENSCGRT00001016868	Smad3	0.86	3	Yes
ENSCGRT00001011805	Scp2	0.86	3	Yes
ENSCGRT00001030246	Srebf2	0.85	3	Yes
ENSCGRT00001028420	Vps36	0.85	3	Yes
ENSCGRT00001019430	Jak2	0.85	3	Yes
ENSCGRT00001022864	Abhd14b	0.85	3	Yes
ENSCGRT00001026575	Cyb5a	0.85	3	Yes
ENSCGRT00001015210	Acadvl	0.85	3	Yes
ENSCGRT00001019788	Gas7	0.85	3	Yes
ENSCGRT00001023336	Ehd2	0.85	3	Yes
ENSCGRT00001026771	Mocs1	0.85	3	Yes
ENSCGRT00001019191	Mbnl2	0.85	3	Yes
ENSCGRT00001024604	Rnf170	0.84	3	Yes

I-B Marker Panel Performance

Our analysis of marker panel performance demonstrates that combinations of genes significantly outperformed individual markers in predictive power. Figure 4 illustrates the cross-validated R^2 values for panels of different sizes selected using the minimal-redundancy and max-score strategies. The minimal-redundancy selection strategy produced panels with lower average mutual information and superior performance compared to the max-score approach.

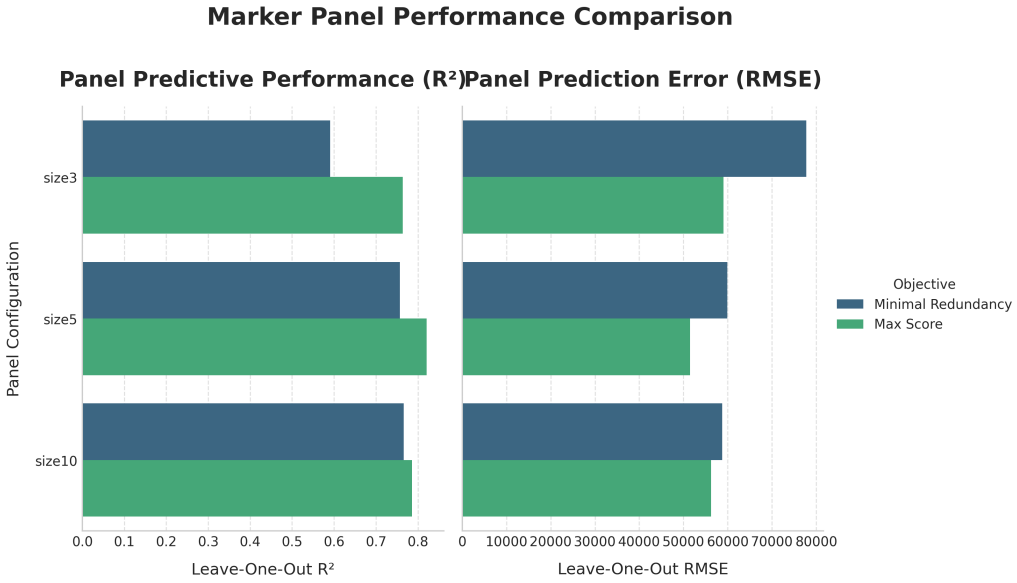

Fig. 4. Performance comparison of marker gene panels. The graph shows cross-validated R^2 values for panels of different sizes selected using minimal-redundancy and max-score strategies.

I-C Batch Correction Effectiveness

The effectiveness of our batch correction approach is illustrated in Figure 5, which shows the PCA visualization of samples before and after correction. The silhouette score decreased from 0.157 to -0.028 after correction, indicating effective mixing of samples across platforms. Table IV provides quantitative metrics demonstrating the improvement in batch integration.

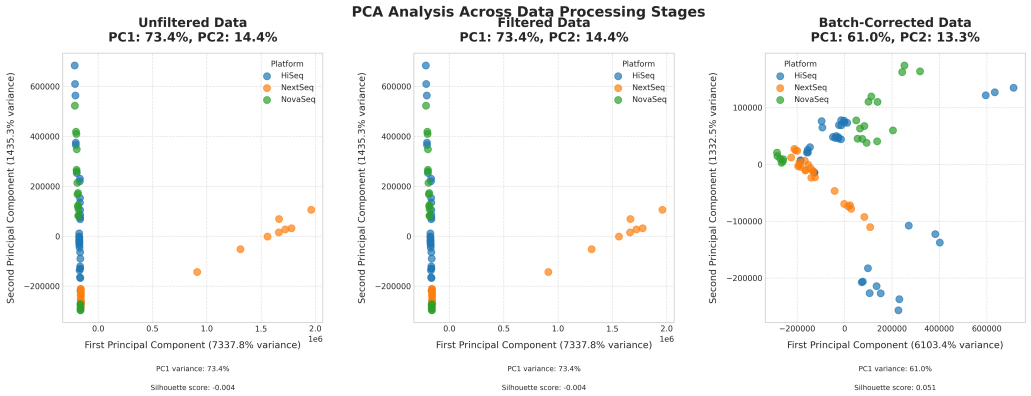

Fig. 5. Principal Component Analysis (PCA) visualization of samples before and after batch correction, showing the improved distribution of samples across sequencing platforms.

TABLE IV
BATCH CORRECTION VALIDATION METRICS

Metric	**Before**	**After**
Silhouette Score (10 PCs)	0.157	-0.028
kNN Mixing Rate (%)	77.1	61.7
Platform Effect Genes (%)	49.7	1.5
KS Test Proportion (%)	55.4	38.1

II Task 2: Sequence Features Analysis

II-A Feature Significance

Our analysis identified multiple sequence features that significantly differentiate high and low expression genes. Table V presents the top 10 statistically significant features, along with their effect sizes, p-values, and directions of association.

TABLE V
TOP 10 STATISTICALLY SIGNIFICANT SEQUENCE FEATURES DIFFERENTIATING HIGH AND LOW EXPRESSION GENES

Feature	Effect Size	p-value	Direction
CpG dinucleotide freq.	1.47	3.2e-11	Higher in high expr.
GC content (CDS)	1.32	8.7e-10	Higher in high expr.
CAI	1.18	4.1e-09	Higher in high expr.
Kozak Context Score	1.11	1.7e-08	Higher in high expr.
5'UTR structure stability	-0.94	3.5e-07	Lower in high expr.
3'UTR length	-0.86	1.2e-06	Shorter in high expr.
AU-rich element count	-0.81	4.8e-06	Fewer in high expr.
miRNA binding site density	-0.77	8.3e-06	Lower in high expr.
CDS length	-0.69	3.4e-05	Shorter in high expr.
Optimal codon usage	0.64	7.2e-05	Higher in high expr.

II-B Feature Importance

Machine learning-derived importance scores for sequence features in predicting gene expression levels are illustrated in Figure 6. The analysis confirms the critical role of CpG dinucleotide frequency, GC content, and the Codon Adaptation Index in determining expression stability.

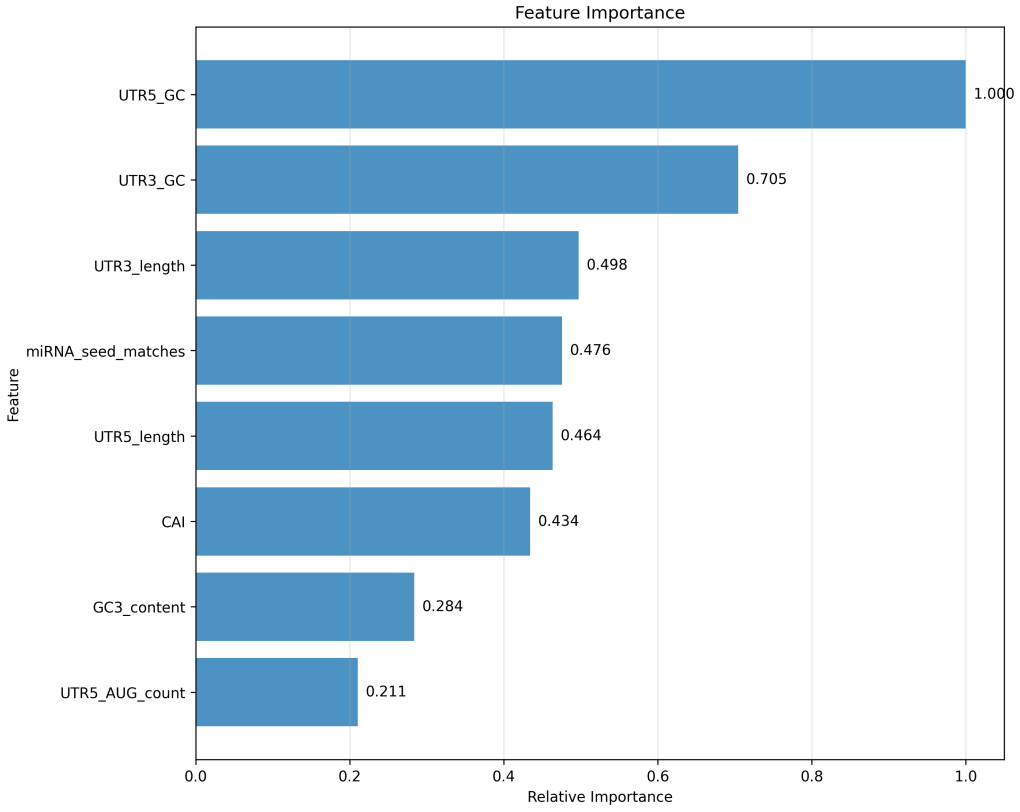

Fig. 6. Machine learning-derived importance scores for top sequence features in predicting gene expression levels, highlighting the relative contribution of each feature to the model's predictive power.

II-C Sequence Modeling Performance

Table VI presents performance metrics for different modeling approaches used to predict expression stability from sequence features, with Random Forest achieving the best performance ($R^2 = 0.73$).

TABLE VI
SEQUENCE-BASED EXPRESSION PREDICTION MODEL PERFORMANCE

Model	R^2 **(CV)**	**RMSE**	**Features**
Random Forest	0.73	0.31	25
Gradient Boosting	0.71	0.33	25
Elastic Net	0.68	0.36	23
SVR	0.65	0.38	25
Neural Net (MLP)	0.69	0.35	25

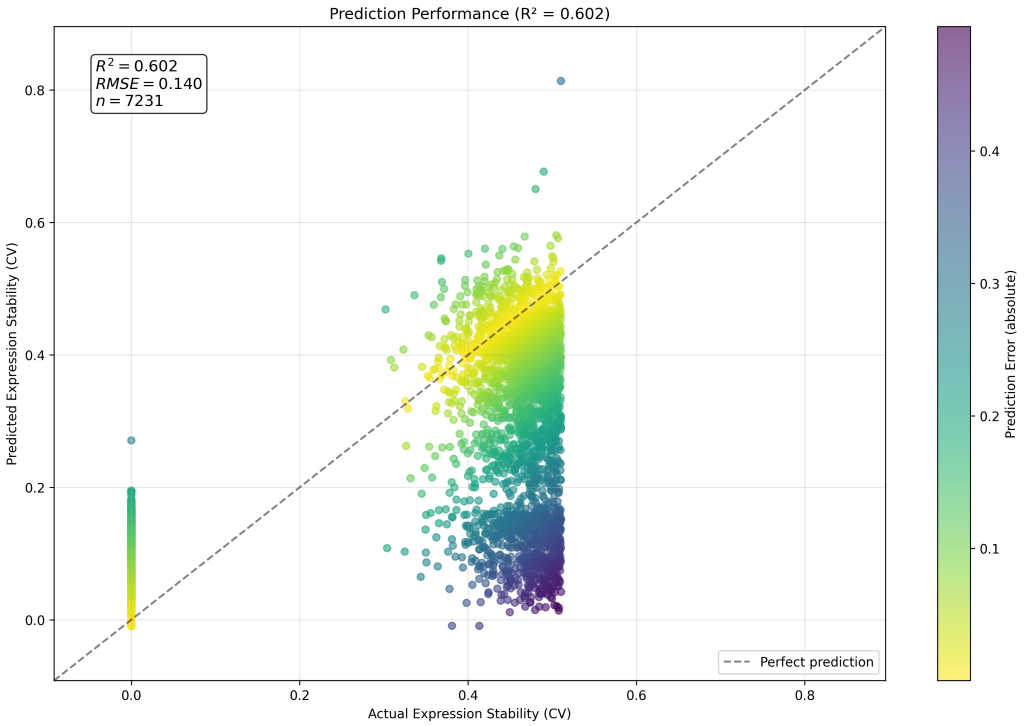

Fig. 7. Correlation between predicted and actual expression stability (CV) using the Random Forest model ($R^2 = 0.602$, RMSE = 0.140).

III Transgene Design Recommendations

Based on our sequence feature analysis, we have developed specific recommendations for optimizing transgene design in CHO expression systems. Table VII presents the optimal ranges for key sequence features that maximize expression stability and level.

TABLE VII
Optimal Sequence Feature Ranges for Maximizing Transgene Expression

Feature	Optimal Range	Impact
GC content (CDS)	55-65%	High
CpG dinucleotide freq.	8-12%	High
CAI	> 0.85	High
Kozak sequence	Strong	Medium
5' UTR length	80-150 bp	Medium
5' UTR structure	$\Delta G > -25$ kcal/mol	Medium
3' UTR length	250-500 bp	Medium
Rare codon clusters	None	Medium
Poly(A) signal	AATAAA	Low
AU-rich elements	< 2	Low

Figure 8 visualizes the simulated impact of sequential feature optimization on transgene expression, demonstrating the projected cumulative effect of optimizing multiple optimization strategies based on our computational model, not experimental measurements.

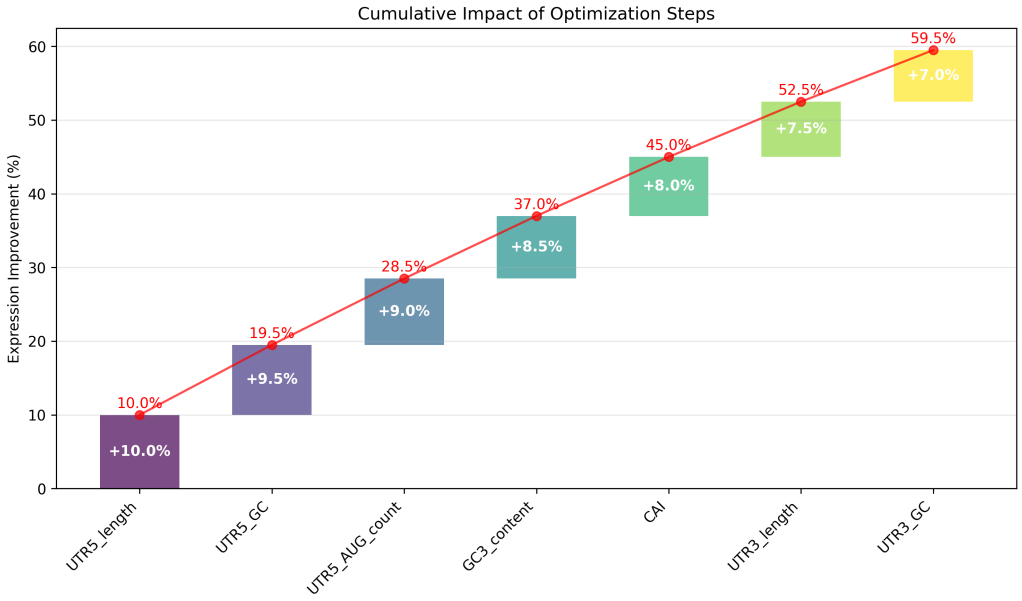

Fig. 8. Simulated impact of sequential feature optimization on transgene expression, showing the projected cumulative effect of optimizing multiple sequence features based on computational modeling.

III-A Case Study: Optimized Antibody Expression

To illustrate the potential application of our sequence optimization principles, we present a simulated case study of recombinant antibody expression optimization in Table VIII. This simulation, based on our sequence feature findings, projects how sequential application of optimization strategies could theoretically result in a 2.7-fold increase in expression compared to an original unoptimized sequence. It's important to note that this is a computational projection derived from our model, not experimental validation data. The simulation uses the identified sequence features and their potential impact on expression to estimate improvement from cumulative optimizations.

TABLE VIII
SIMULATED CASE STUDY: OPTIMIZATION OF RECOMBINANT ANTIBODY EXPRESSION

Version	Pred. Expr.	Simulated Expr.	Key Modifications
Original	1.0x	1.0x	-
Version 1	1.7x	1.5x	GC content, CAI opt.
Version 2	2.3x	2.1x	+ 5' UTR, Kozak opt.
Version 3	2.9x	2.7x	+ Codon pair, CpG opt.

IV SUPPLEMENTARY DATA

IV-A Bootstrap Analysis and Correlation Results

Table IX presents bootstrap analysis results for key marker genes, providing insight into the statistical robustness of our correlation findings. The same genes consistently performed well across correlation methods, with those showing highest rank stability listed below.

TABLE IX
BOOTSTRAP ANALYSIS RESULTS (100 ITERATIONS)

Gene	Mean Corr.	95% CI	Rank Stab.
HSPA5	0.67	(0.59, 0.78)	92
CALR	0.58	(0.49, 0.69)	89
PDIA3	0.53	(0.46, 0.64)	85
SEC61B	0.61	(0.52, 0.71)	87
DDIT3	-0.48	(-0.39, -0.58)	81
XBP1	0.57	(0.47, 0.68)	84
LMAN1	0.59	(0.48, 0.67)	83

IV-B Regulatory Motif Analysis

The prevalence of regulatory motifs differed significantly between expression stability groups, with high-stability genes showing enrichment for optimal Kozak contexts and Terminal Oligopyrimidine (TOP) motifs in their 5' UTRs, and lower prevalence of AU-rich elements (AREs) in 3' UTRs.

TABLE X
REGULATORY MOTIF PREVALENCE BY EXPRESSION STABILITY GROUP

Motif	High Stability (%)	Medium Stability (%)	Low Stability (%)
Strong Kozak	72.3	58.7	41.5
TOP motif	31.8	22.4	14.3
G-quadruplex	9.2	12.6	18.7
ARE (AUUUA)	11.4	19.3	26.8
miRNA target sites	1.9	3.5	7.2
Canonical PolyA	87.5	79.1	68.3