A NOVEL AI/ML APPROACH TO CODON OPTIMIZATION FOR ENHANCED ANTIBODY EXPRESSION

Andrew Tsao, Dashan Sun, Deng Feng, Hui Li, Zhifei Sun, Gaoxu Xue, Crystal Richardson, Emily Sutton, Laura Kelly, Riley Graham, Sumit Kumar, Guangyou Duan, Ethan Ge

GENEWIZ from Azenta Life Sciences, South Plainfield, NJ 07080

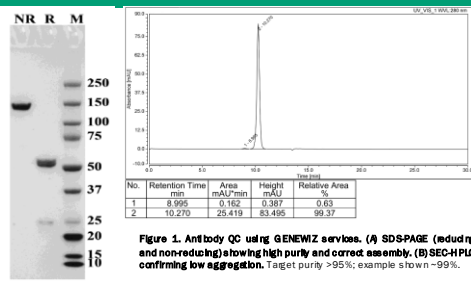
Abstract

In the rapidly advancing field of antibody discovery - driven by AI and machine learning - achieving high and consistent protein expression remains a critical challenge. Traditional codon optimization methods rely on static metrics such as codon adaptation index (CAI), GC content, and minimum free energy (MFE), which fail to capture the complex, non-linear relationship between sequence design and expression outcomes. To address this, we developed an AI-driven codon optimization platform that combines a proven heuristic framework with deep learning trained on large-scale genomic and proprietary high-expression datasets. This approach captures real expression biology by learning non-linear sequence–expression relationships and identifying sequence patterns associated with high protein output. Our platform outperforms conventional codon optimization methods, delivering higher antibody yields and more consistent expression across diverse constructs, as demonstrated by wet-lab validation.

Materials & Methods

The GENEWIZ AI/ML-powered codon optimization model was developed using large-scale genomic sequence data from proprietary in-house expression datasets, and publicly available high-expression sequence datasets. The model was trained to capture non-linear sequence features associated with protein expression and further verified using experimentally validated high-expression constructs. For benchmarking, a panel of antibody sequences (wild-type variable regions) was selected and independently optimized using the GENEWIZ AI-based approach alongside multiple commercially available codon optimization tools. Each optimized sequence was synthesized, cloned into standardized expression vectors, and expressed under consistent conditions to enable direct comparison.

Recombinant antibodies were produced via transient expression in HEK293 or CHO systems across multiple scales (1 mL, 3 mL, and 10 mL for screening). Expression performance was evaluated based on yield and consistency across constructs, corresponding to the datasets presented in the figures. Purified antibodies were obtained using affinity chromatography (e.g., Protein A), with additional polishing steps applied as needed. Product quality was assessed by A280 quantification, SDS-PAGE (reduced and non-reduced), and SEC-HPLC for aggregation analysis. Endotoxin levels were measured using LAL assay (<1 EU/mg), and overall purity exceeded 95% across samples.



No.	Retention Time min	Area mAU*min	Height mAU	Relative Area %
1	8.995	0.162	0.387	0.63
2	10.270	25.419	83.495	99.37

Figure 1. Antibody QC using GENEWIZ services. (A) SDS-PAGE (reducing and non-reducing) showing high purity and correct assembly. (B) SEC-HPLC confirming low aggregation. Target purity >95%; example shown ~99%.

Results

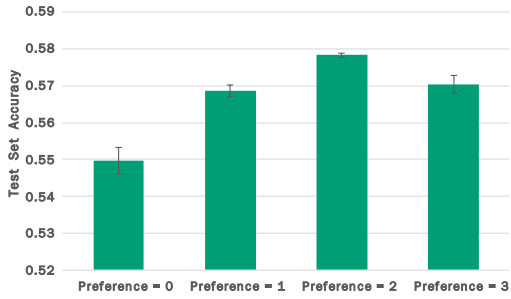


Figure 2. Higher-expression training data improves model performance on held-out test sequences

Models trained on higher-expression datasets showed improved performance on held-out test sequences. Each bar represents expression-prediction accuracy after training with a different Preference-level dataset, where higher preference indicates higher protein expression in the training data. Test accuracy increased from Preference 0 to 2, showing that the model learned meaningful sequence features linked to improved expression and could apply them to unseen sequences.

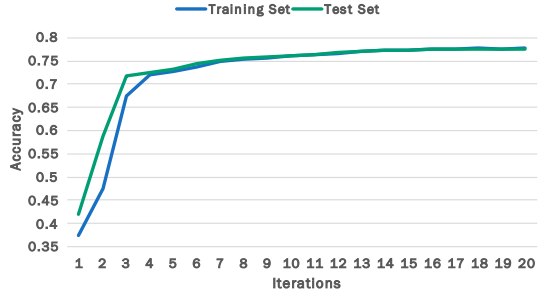


Figure 3. AI/ML model performance improves over training while maintaining generalization

Training and test-set accuracy were monitored during model training process. Accuracy increased rapidly during early iterations and then plateaued, with training and test-set performance remaining closely aligned. This pattern indicates that the model learned predictive sequence-expression features while maintaining generalization to held-out test data, rather than relying only on memorization of the training set.

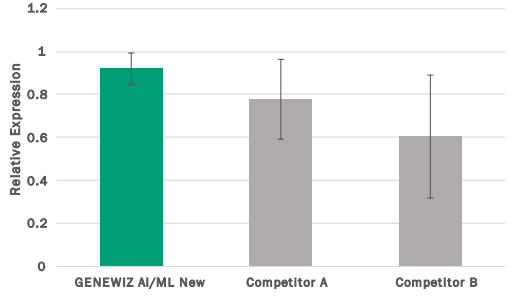


Figure 4. GENEWIZ codon optimization improves average antibody expression compared with other commercial tools

A panel of antibody variable-region sequences was optimized using the GENEWIZ AI-powered codon optimization model and two commercially available codon optimization tools. Optimized sequences were synthesized, cloned, and expressed under comparable conditions. The GENEWIZ AI-powered approach showed higher average relative expression compared with Competitor A and Competitor B. Error bars represent variability across tested constructs.

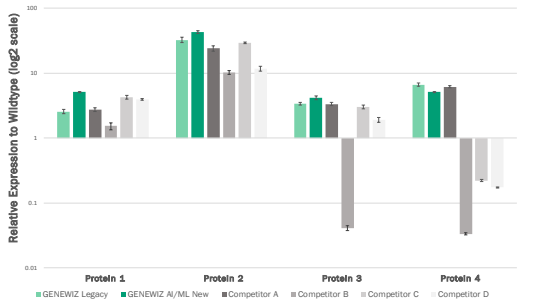


Figure 5. Higher protein expression with GENEWIZ codon optimization tool.

Protein expression levels were quantified and normalized to the corresponding wild-type (WT) expression for each protein. Data are presented as $\log_2$-transformed fold change relative to WT. Codon optimization was performed using our tools and four competitor methods. For each protein–tool combination, expression was measured using three biological replicates, each with three technical replicates.

Conclusions

- Reliable, consistent optimization across genes and constructs - built to generalize beyond training data and deliver dependable codon optimization for diverse sequence designs.
- Higher protein and antibody expression: GENEWIZ AI/ML-powered optimization benchmarked against comparable tools to help improve expression and maximize protein yield.
- Built directly into our Gene Synthesis ordering form: customers can design, optimize, and order constructs in one streamlined workflow.

Scan QR Code for more detailed information