absci | PEGS BOSTON 2023

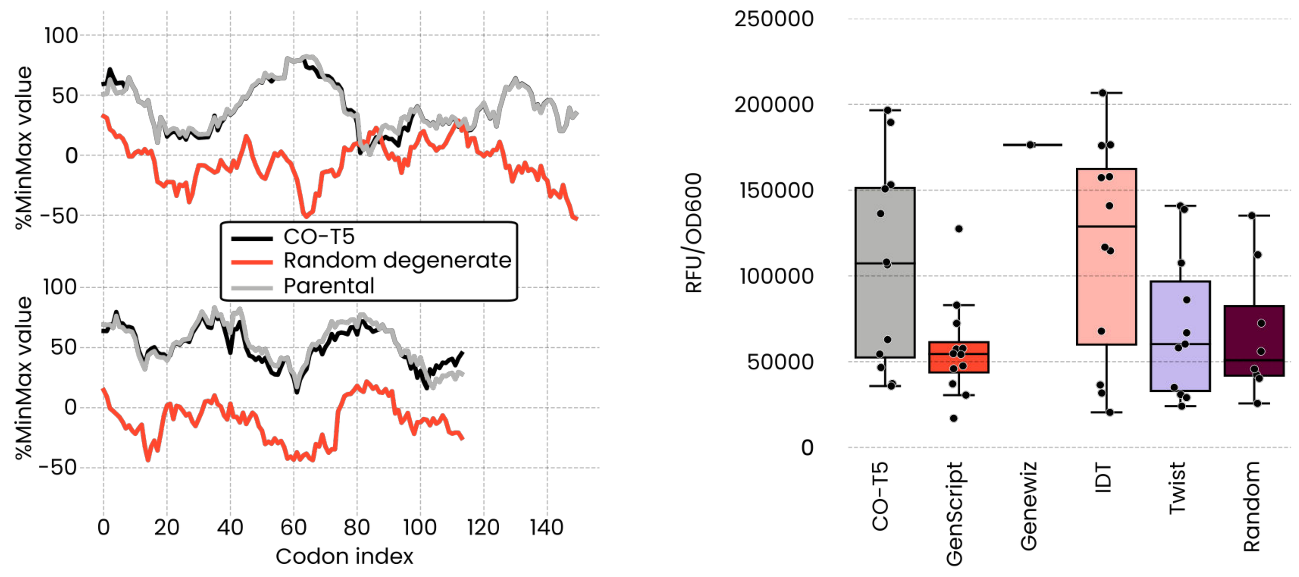
DEEP LEARNING-BASED CODON OPTIMIZATION WITH LARGE-SCALE SYNONYMOUS VARIANT DATASETS ENABLES GENERALIZED TUNABLE PROTEIN EXPRESSION

David A. Constant, **Jahir M. Gutierrez**, Anand V. Sastry, Rebecca Viazzo , Nicholas R. Smith, Jubair Hossain, **David A. Spencer**, Hayley Carter, Abigail B. Ventura, Michael T. M. Louie, Christa Kohnert, Rebecca Consbruck, Joshua Bennett, Kenneth A. Crawford, John M. Sutton, Anneliese Morrison, Andrea K. Steiger, Kerianne A. Jackson, Jennifer T. Stanton, Shaheed Abdulhaqq, Gregory Hannum, Joshua Meier, Matthew Weinstock, Miles Gander

ABSTRACT

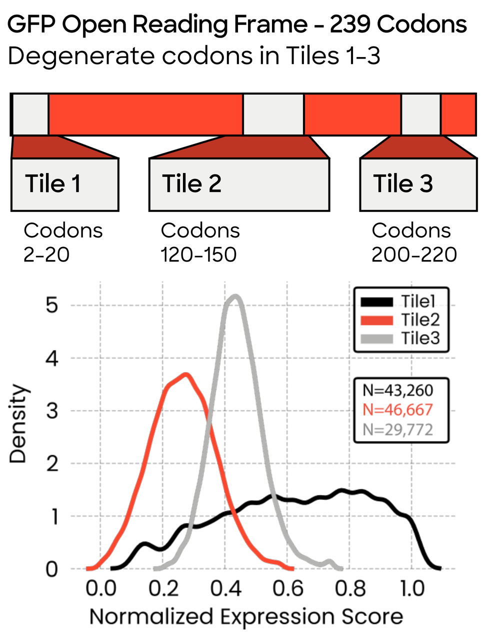
Increasing recombinant protein expression is of broad interest in industrial biotechnology, synthetic biology, and basic research. Codon optimization is an important step in heterologous gene expression that can have dramatic effects on protein expression level. Several codon optimization strategies have been developed to enhance expression, but these are largely based on bulk usage of highly frequent codons in the host genome and can produce unreliable results. Here, we develop deep contextual language models that learn the codon usage rules from natural protein coding sequences across members of the *Enterobacterales* order. We then fine-tune these models with over 150,000 functional expression measurements of synonymous coding sequences from three proteins to predict expression in *E. coli*. We find that our models recapitulate natural context specific patterns of codon usage and can accurately predict expression levels across synonymous sequences. Finally, we show that expression predictions can generalize across proteins unseen during training, allowing for *in silico* design of gene sequences for optimal expression. Our approach provides a novel and reliable method for tuning gene expression with many potential applications in biotechnology and biomanufacturing.

SEQUENCES OPTIMIZED FOR NATURAL CODON USAGE HAVE VARIABLE EXPRESSION



Left: Two representative %MinMax codon usage profiles of CDSs in the holdout test set. The comparison is made across parental natural sequence (grey), CO-T5-model-generated sequence (black), and sequence of randomly sampled synonymous codons (red). **Right:** Normalized fluorescence values for CO-T5-generated GFP variants, sequences optimized with commercial tools, and random synonymous sequences.

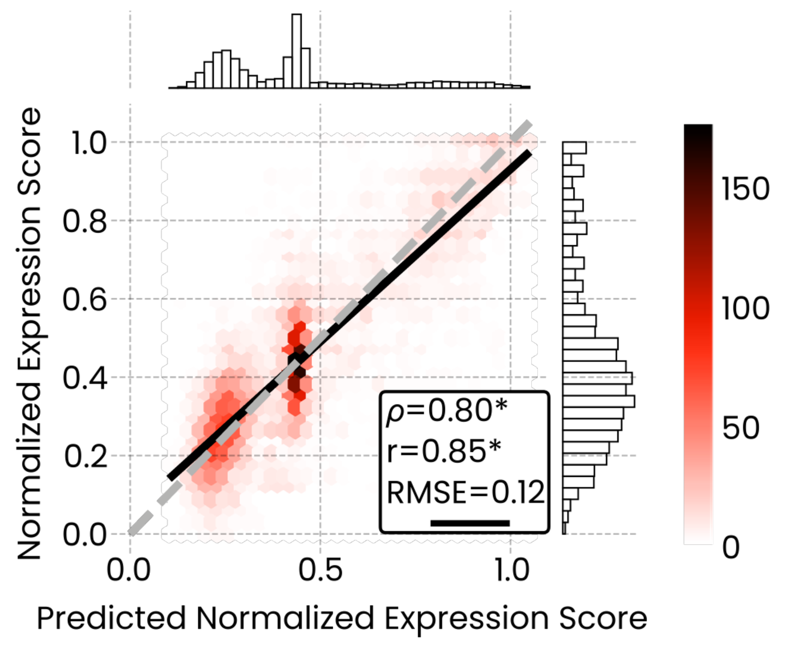
MODEL PREDICTED EXPRESSION LEVELS OF GFP LIBRARIES



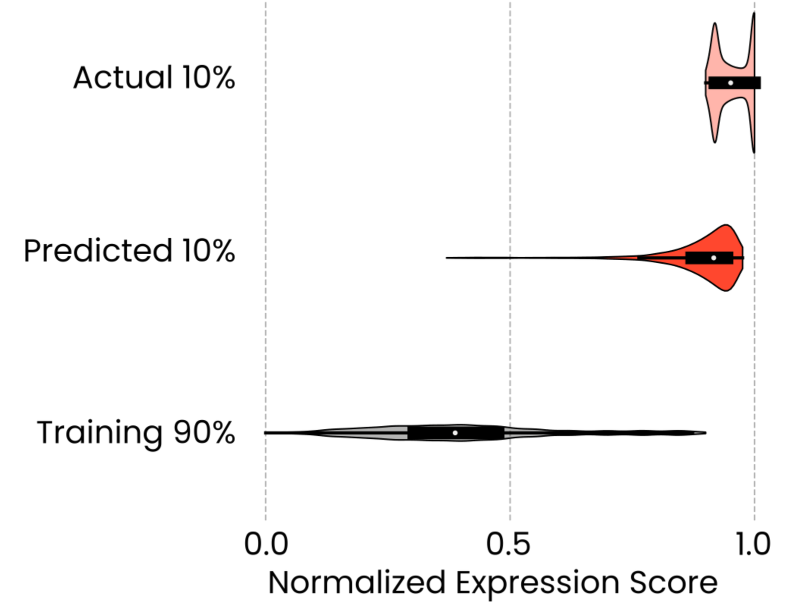
Left: (Top) Three tile synonymous mutant libraries constructed for GFP. (Bottom) Density plots of normalized expression scores for all three tile libraries. Distributions show varied expression score profiles based on tile position in CDS.

Right: (A) Heatmap correlation of GFP model predictions of a test set against measured expression scores. Dashed line represents unity; solid line is best fit. (B) Violin plots depicting model performance of out-of-distribution expression predictions for a holdout set consisting of the top 10% expressing GFP variants.

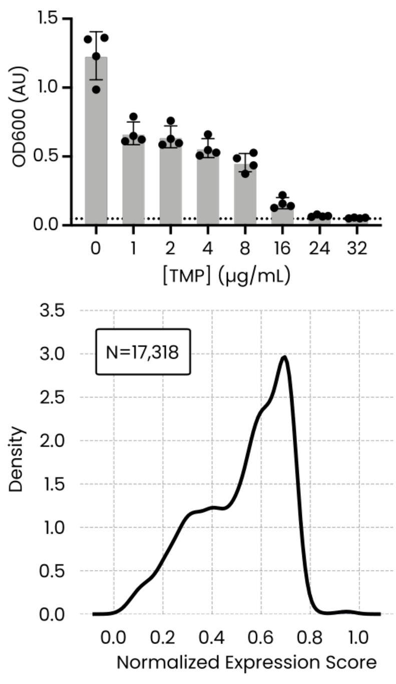
A



B



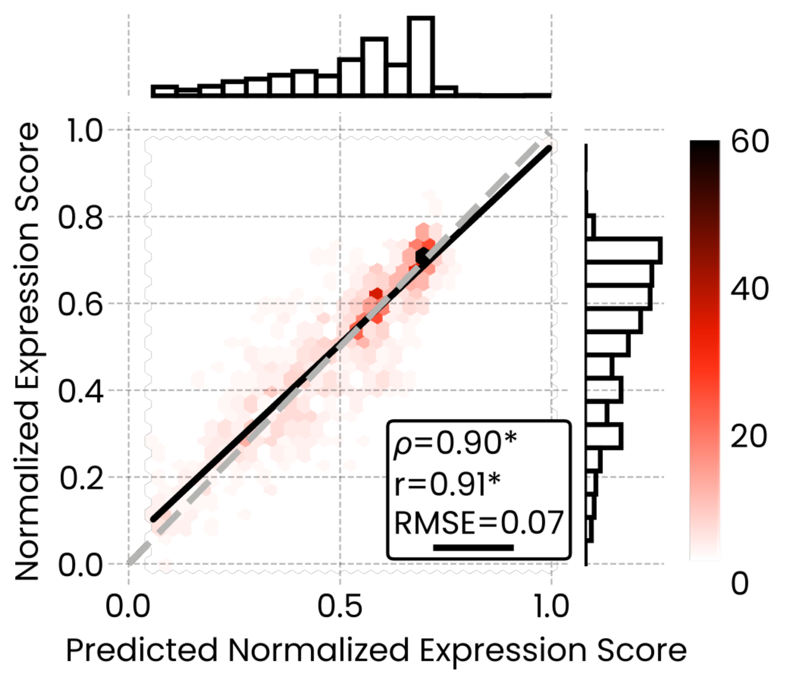
MODEL PREDICTED EXPRESSION LEVELS OF FOLA LIBRARIES



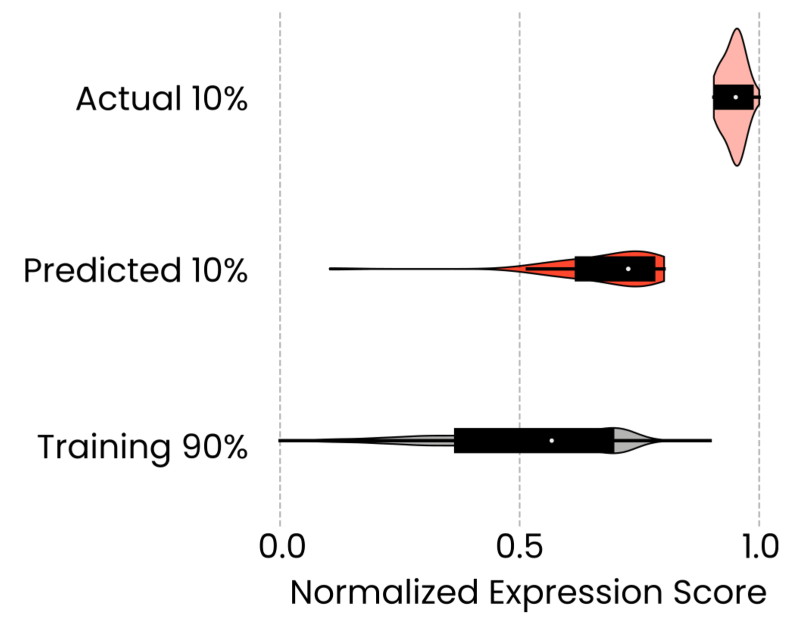
Left: (Top) OD measurements of folA degenerate libraries after 24hr of growth at increasing levels of TMP. (Bottom) Normalized expression score distribution from folA library selections.

Right: (A) Heat map correlation of model-predicted expression score values for folA variants and expression score measurements for a holdout test set. Dashed line represents unity; solid line is best fit. (B) Violin plots depicting model performance of out-of-distribution expression predictions for a holdout set consisting of the top 10% expressing folA variants.

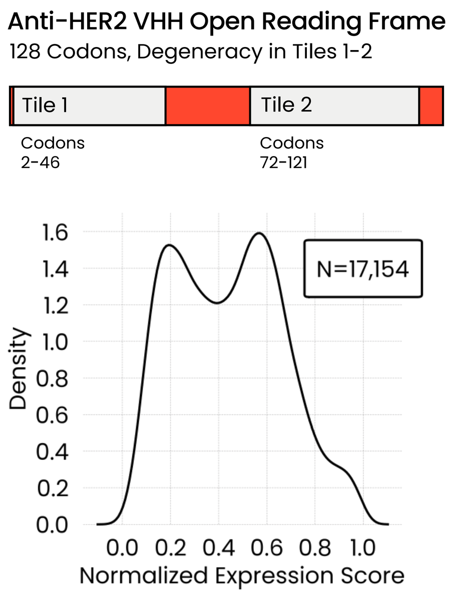
A



B



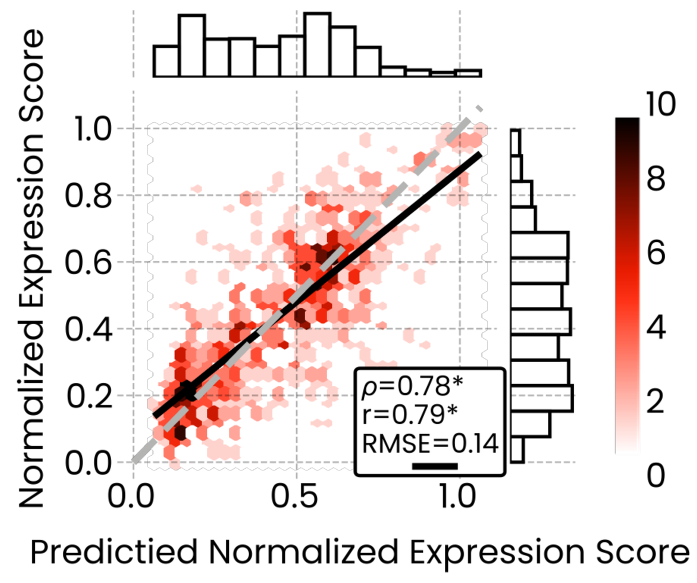
MODEL PREDICTED EXPRESSION LEVELS OF VHH LIBRARIES



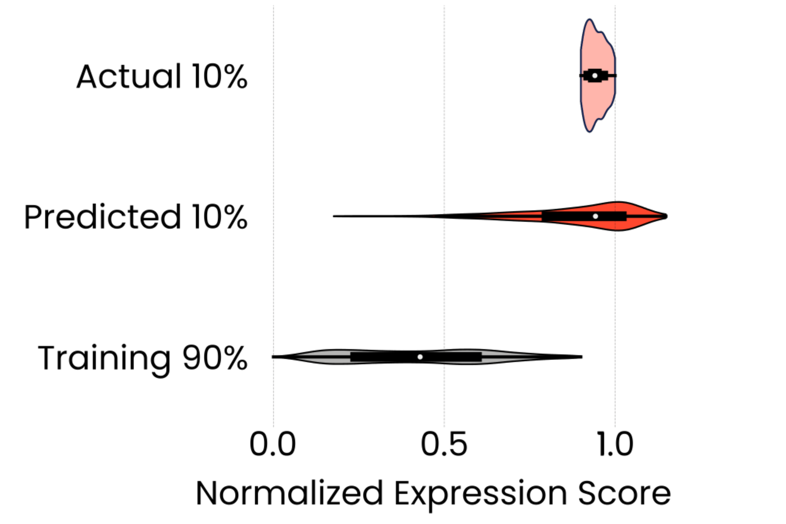
Left: (Top) Schematic of degenerate tiles for the anti-HER2 VHH parent sequence. (Bottom) Normalized expression score distribution of VHH library variants.

Right: (A) Heatmap correlation of model-predicted VHH expression scores and expression score measurements. Dashed line represents unity; solid line is best fit. (B) Violin plots depicting model performance of out-of-distribution expression predictions for holdout set consisting of top 10% expressing anti-HER2 VHH variants.

A



B



MULTI-PROTEIN LEARNING

Test dataset	Training dataset	Spearman ρ	Pearson r
GFP	GFP	0.759	0.808
	GFP+*folA*	0.776	0.830
	GFP+VHH	0.766	0.824
	ALL	**0.788**	**0.856**
folA	*folA*	0.899	0.907
	GFP+*folA*	0.916	0.910
	folA+VHH	0.905	0.900
	ALL	**0.918**	**0.919**
VHH	VHH	0.782	0.794
	GFP+VHH	0.804	0.815
	folA+VHH	**0.837**	**0.848**
	ALL	0.818	0.824

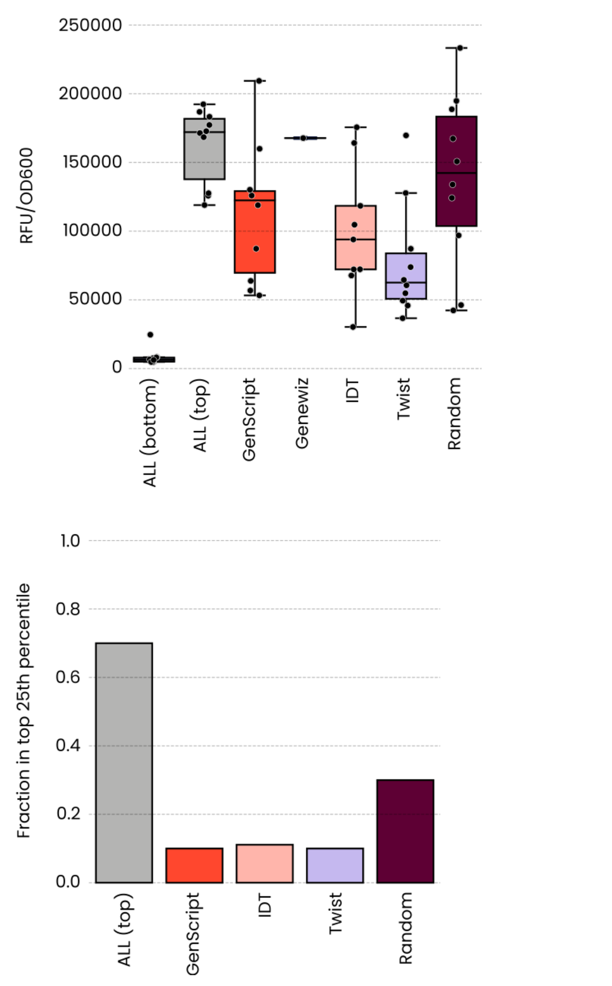
Performance of fine-tuned CO-BERTa models across the GFP (top), folA (middle), and anti-HER2 VHH (bottom) holdout datasets. In all cases we observed increased performance when training on multiple proteins. All p-values across Spearman and Pearson correlations are significant (p<0.01).

TUNING EXPRESSION WITH MODEL-DESIGNED DNA VARIANTS FOR UNSEEN PROTEINS

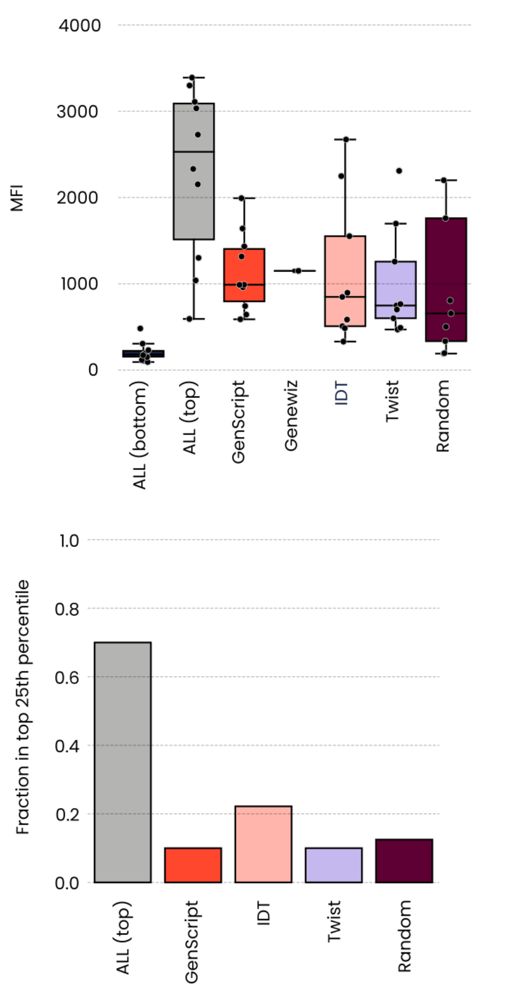
Top: Normalized expression as measured by RFU/OD600 for mCherry (left) or mean fluorescent intensity (MFI) for anti-SARS-CoV-2 VHH (right) for sequences designed by the ALL model to maximize or minimize expression, compared to various optimization baselines.

Bottom: Barcharts showing fraction of designs in the upper quartile of all measured mCherry (left) or anti-SARS-CoV-2 VHH (right) variants for each sequences group, excluding the ALL (bottom) set. Genewiz condition single sequence excluded.

mCherry



anti-SARS-CoV-2 VHH



PARTNER WITH US

Leverage our deep-learning based codon optimization technology to enable your projects.

absci PEGS SUMMIT