Article

Not peer-reviewed version

Deep Learning-Guided Reverse Translation Enhances Soluble Expression of Recombinant Proteins in *Escherichia coli*

Dong Yu , Nan Geng , Lin Fan , Yanmei Qin , Shangshang Sun , Hao Chen , Rouyu Wang , Xiaoping Liao * , Chun You *

Posted Date: 15 May 2026

doi: 10.20944/preprints202605.1014.v1

Keywords: deep learning; *Escherichia coli*; recombinant protein; codon optimization; soluble expression

 doi:10.20944/preprints202605.1014.v1

Article

Deep Learning-Guided Reverse Translation Enhances Soluble Expression of Recombinant Proteins in *Escherichia coli*

Dong Yu [1], Nan Geng [1], Lin Fan [2,3], Yanmei Qin [3,4], Shangshang Sun [3], Hao Chen [1], Ruoyu Wang [3], Xiaoping Liao [3,*] and Chun You [3,4,5,*]

[1] Tianjin University of Science and Technology, Tianjin 300457, China
[2] Sino-Danish College, University of Chinese Academy of Sciences, Beijing, China
[3] Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China
[4] University of Chinese Academy of Sciences, Beijing 100049, China
[5] State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China
* Correspondence: liao_xp@tib.cas.cn (X.L.); yoshion@foxmail.com (C.Y.)

Abstract

Enhancing the soluble expression of heterologous proteins in chassis microorganisms is critical for fundamental biological research and synthetic biology-driven industrial applications. Current methods for designing DNA sequences to ensure high soluble expression often rely excessively on high-frequency codons while overlooking optimal codon context, leading to suboptimal outcomes. To address these limitations, we developed an integrated deep learning framework combining a synonymous codon generation (SCG) model and a gene expression level prediction (GELP) model. The SCG model captures codon usage patterns in *Escherichia coli* using large-scale genomic data, whereas the GELP model leverages gene expression data to prioritize sequences with high soluble expression potential. We validated our approach by optimizing the DNA sequences of two industrial enzymes, α-glucan phosphorylase (αGP) and isoamylase (IA), achieving 20.52-fold and 3.05-fold increases in soluble expression, respectively, relative to the wild type. This study provides a powerful tool for designing DNA sequences that confer high soluble expression and for understanding the relationship between DNA sequence and protein expression. Notably, SCG-GELP reveals a protein surface-targeted codon optimization strategy that substantially enhances soluble protein yield. The framework is publicly accessible at https://scg-gelp.biodesign.ac.cn, and its open-source code and trained models are available on GitHub at https://github.com/yuddecho/SCG-GELP.

Keywords: deep learning; *Escherichia coli*; recombinant protein; codon optimization; soluble expression

1. Introduction

Enzyme cost critically determines the economic viability of enzymatic processes at industrial scales. Established cost-reduction strategies include protein engineering to improve enzyme thermostability and activity [1], enzyme immobilization [2], and optimizing soluble protein expression [3]. Among these approaches, optimizing soluble protein expression permits straightforward monitoring via sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), providing a rapid and cost-effective screening method. Consequently, enhancing the soluble expression of heterologous recombinant proteins in *E. coli* has emerged as a key focus area with substantial scientific and industrial relevance.

The soluble expression of heterologous recombinant proteins is influenced by multiple factors, including codon usage bias, host metabolic capacity, and culture conditions. Although expression

 doi:10.20944/preprints202605.1014.v1

yields can be improved by optimizing culture conditions [4,5] —such as adjusting medium composition, lowering incubation temperature, controlling expression rates with weak promoters, and co-expressing molecular chaperones [6] —the intrinsic DNA sequence remains a primary determinant of expression efficiency. It plays a pivotal role in regulating mRNA stability, translation kinetics, and soluble protein yield [7,8], and sequence optimization therefore exerts a greater impact on expression levels than external condition adjustments.

The process of generating a DNA sequence from a protein sequence is termed reverse translation. Because the genetic code is degenerate, a single protein sequence can be encoded by a vast number of potential DNA sequences [9]. To identify optimized DNA sequences that enhance heterologous soluble expression, codon optimization is employed, taking into account various factors such as codon usage bias, codon pair context, tRNA availability, GC content, ribosome binding sites (RBS), hidden termination codons, motif avoidance, restriction site removal, mRNA secondary structure, and hydrophilicity index. Numerous computational tools, including DNAWorks [10], JCat [11], OPTIMIZER [12], mRNA Optimizer [13], Gene Designer [14], Visual Gene Developer [14], COOL [15], and D-Tailor [16], have been developed to facilitate codon optimization by adjusting these factors. However, owing to the limited understanding of host cellular mechanisms, the expression levels of genes optimized using these traditional methods often fall short of expectations [17]. Recently, deep learning approaches have outperformed traditional methods in codon optimization by capturing complex nonlinear relationships and hidden patterns among expression-influencing factors, offering a promising alternative [18–20]. Accordingly, several deep learning-based codon optimization methods have been developed, including BiLSTM-CRF [21], ICOR [22], CodonBERT [23], CodonTransformer [24], and DeepCodon [25]. Nevertheless, the BiLSTM-CRF model was trained on a small dataset and fails to capture the relationship between DNA sequences and expression levels. ICOR lacks experimental validation and does not adequately model the sequence-expression relationship. CodonBERT, despite its advanced design, is limited by a small fine-tuning dataset comprising only three proteins and relies on normalized fluorescence values rather than actual expression levels, compromising its generalizability to industrial proteins.

To overcome these limitations, we present SCG-GELP, a novel deep learning framework that integrates de novo DNA sequence generation with soluble protein expression prediction, enabling robust expression optimization in *E. coli*. The SCG model employs a transformer-based encoder-decoder architecture trained on *E. coli* genome data to generate synonymous codon sequences that comply with host-specific usage patterns, whereas the GELP model combines support vector machine (SVM) [26], multi-layer perceptron (MLP), and logistic regression (LR) algorithms with multimodal input features to identify sequences predisposed to high soluble expression. This dual-model approach ensures that optimized sequences not only adhere to host-specific codon usage patterns but also encode expression-enhancing features overlooked by rule-based tools. We rigorously validated SCG-GELP through experimental testing of two industrially relevant enzymes. For α-glucan phosphorylase (αGP), the optimized sequence (αGP-Opt2) exhibited a 15.55-fold increase in total expression in whole-cell lysates and a 20.52-fold increase in soluble yield compared to the wild-type sequence, outperforming the GenScript-optimized design by 5.76-fold and 6.78-fold, respectively. Similarly, the isoamylase (IA) variant (IA-Opt1) achieved a soluble protein fraction of 17.5% in supernatants—nearly double that of the GenScript-optimized sequence (9.0%). These results demonstrate SCG-GELP's ability to uncover non-obvious sequence determinants of expression while maintaining translational fidelity, offering an effective alternative to conventional codon optimization strategies.

2. Results

2.1. Overview of the SCG-GELP Framework

The SCG-GELP framework integrates two collaborative models to systematically optimize DNA sequences for high soluble expression in *E. coli* (Figure 1). The SCG model uses a Transformer-based

 doi:10.20944/preprints202605.1014.v1

encoder-decoder architecture to reverse-translate the input protein sequence into numerous high-quality candidate DNA variants. The encoder processes the protein sequence through amino acid embedding and multi-head self-attention layers to extract hierarchical features, while the decoder autoregressively generates codon sequences consistent with *E. coli* codon usage patterns. The generated DNA sequences are subsequently evaluated by the GELP model, which extracts sequence features through fine-tuned DNABERT-2 [27] and feeds them into an ensemble classifier (SVM, MLP, and LR). The sequence with the highest mean predicted probability of conferring high soluble expression is selected as the final output.

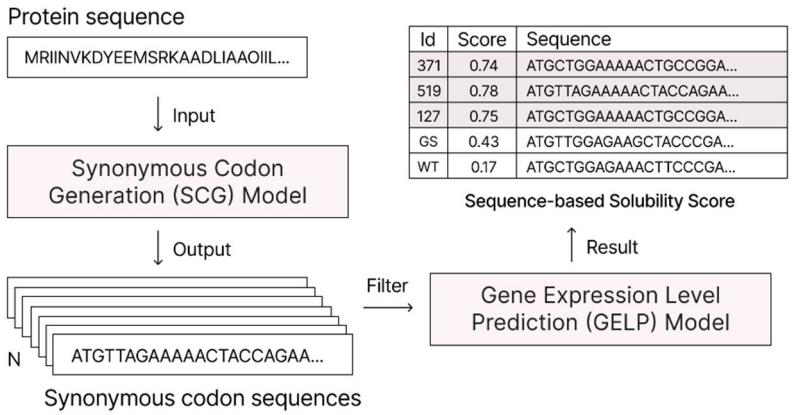

Figure 1. Overview of the SCG-GELP framework. Starting from a protein sequence, the SCG model generates candidate DNA sequences, which are then screened by the GELP model to select the sequence with the highest predicted soluble expression.

2.2. SCG Model Performance Evaluation

We trained the SCG model on the RefSeqE (Reference Sequence *E. coli*) dataset (98,855 protein-DNA sequence pairs). The training loss decreased steadily throughout optimization, whereas the test loss dropped rapidly during the first 10 epochs and subsequently plateaued at approximately 0.97–0.98. Using early stopping, the final test loss converged to 0.969.

To evaluate the performance of the SCG model, we employed a decoding strategy integrating beam search and pruning operations to generate large batches of synonymous codon sequences for proteins in the test set. The top-scoring predicted sequences were selected, and the Codon Adaptation Index (CAI) [28] and guanine–cytosine (GC) content were calculated and compared against the corresponding real sequences in the test set, as shown in Figure 2. Figure 2a presents the CAI distribution between real and predicted sequences, with the horizontal axis representing real sequence values, the vertical axis representing predicted sequence values, the dashed line indicating the diagonal y = x, and the red line showing the fitted curve for the scatter plot; kernel density histograms for the real and predicted sequence CAIs are displayed at the top and right of the panel, respectively. Figure 2b shows the corresponding GC content distribution, using the same layout. For the CAI distribution, we observed that the CAI values of most sequences were improved after optimization, indicating that the model tends to favor codons with higher usage preference during sequence design. Moreover, the peak CAI value of the optimized sequences remained approximately 0.9, suggesting that while the model preferentially selects high-frequency codons, it does not extremize all codons to the most preferred option; instead, it generates rare codons that retain functional value [29]. Regarding GC content (Figure 2b), sequences with originally low GC content tended to be further reduced after optimization, whereas those with originally high GC content tended to increase, producing two distinct peaks in the post-optimization density histogram. The GC content of most optimized sequences fell within the range of 0.25 to 0.7. Collectively, the CAI and GC

 doi:10.20944/preprints202605.1014.v1

content distributions demonstrate that the SCG model effectively elevates CAI and optimizes GC content during synonymous codon generation.

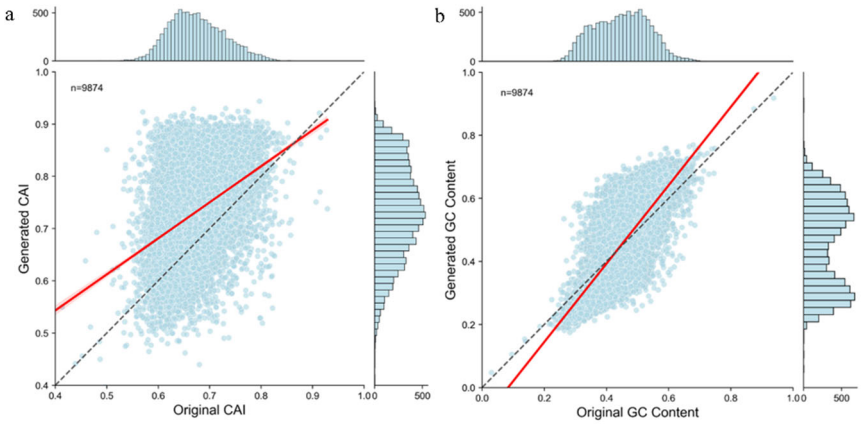

Figure 2. Performance metrics of the SCG model. (a) Codon Adaptation Index (CAI) distribution plot. (b) Distribution plot of guanine-cytosine (GC) content. The horizontal axis is the real sequence values, the vertical axis is the predicted sequence values, the dashed line is the coordinate axis angular bisector y = x, the red line is the scatter fit curve on the graph, and the histograms of kernel density statistics for the real and predicted sequences are shown at the top and right of the graph, respectively.

2.3. GELP Model Performance Evaluation

The GELP model integrates DNABERT-2 for feature extraction and three machine learning classifiers (SVM, MLP, and LR) to predict whether a DNA sequence will yield high soluble expression. We fine-tuned DNABERT-2 on the NESG-DNA (Northeast Structural Genomics) dataset using default training parameters; after 1,400 training steps, the model achieved a test loss of 0.5868 and an accuracy of 70.79%. In addition, we evaluated multiple performance metrics for each of the three classifiers using ten-fold cross-validation on the NESG-DNA (Northeast Structural Genomics) dataset. The average accuracy across all three algorithms exceeded 75%, with SVM achieving the highest accuracy at 81.17%.

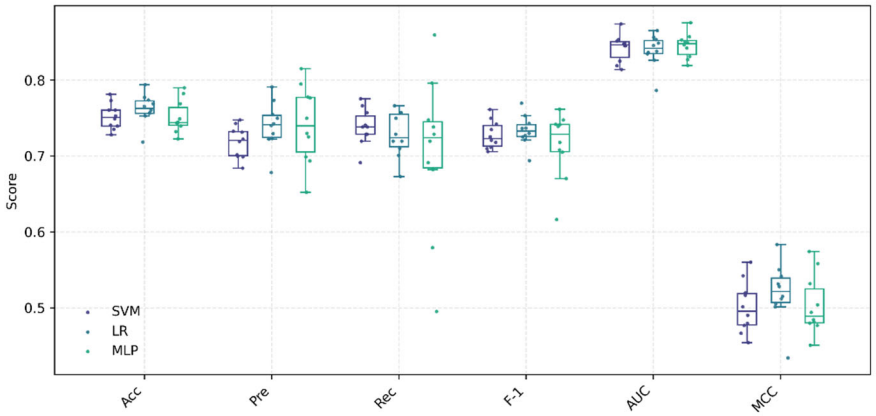

Figure 3. Ten-fold cross-validation performance of the SVM, MLP, and LR classifiers on the NESG-DNA dataset.

2.4. Experimental Verification Results

To rigorously evaluate the SCG-GELP framework, we conducted systematic expression tests using two industrially relevant enzymes: α-glucan phosphorylase (αGP) from Thermotoga maritima

 doi:10.20944/preprints202605.1014.v1

and isoamylase (IA) from Sulfolobus tokodaii (see Supplementary Table S4 for detailed protein information). αGP catalyzes the reversible phosphorylation of α-glucan and serves as a key enzyme in numerous in vitro synthetic enzymatic biosystems that utilize α-glucan to produce hydrogen [30], electricity [31], and inositol [32]. IA hydrolyzes α-1,6-glucosidic branch linkages in glycogen and amylopectin, yielding amylodextrin for complete utilization of branched α-glucan [33]. The experimental design compared protein expression levels among wild-type sequences, GenScript-optimized sequences, and our SCG-GELP-optimized variants (Opt-1 and Opt-2) in an *E. coli* BL21(DE3) expression system.

For αGP, SDS-PAGE quantification revealed striking differences among the tested constructs (Figure 4a). The wild-type (WT) sequence showed minimal detectable expression (5.6% in whole-cell lysate, 4.4% in supernatant), whereas the optimized variant Opt-2 achieved 49.8% expression in lysate and 47.7% in supernatant. This corresponds to a 15.55-fold improvement in total expression and a 20.52-fold increase in soluble fraction relative to wild-type. Notably, Opt-2 outperformed the commercial GenScript-optimized sequence by 5.76-fold in lysate and 6.78-fold in supernatant, while achieving markedly higher soluble protein yields.

The IA expression results presented a more nuanced scenario (Figure 4b). Both the GenScript-optimized and SCG-GELP-optimized sequences showed similarly high total expression levels (31.8% and 26.6% in lysate, respectively), representing approximately 8.3-fold and 6.7-fold improvements over wild-type, respectively. However, crucial differences emerged in soluble expression. IA-Opt1 achieved a 3.05-fold increase in soluble expression compared to WT, with a soluble protein fraction of 17.5% in supernatants—nearly double that of the GenScript-optimized sequence (9.0%). This demonstrates that SCG-GELP optimizes not only total expression but also proper protein folding and soluble expression.

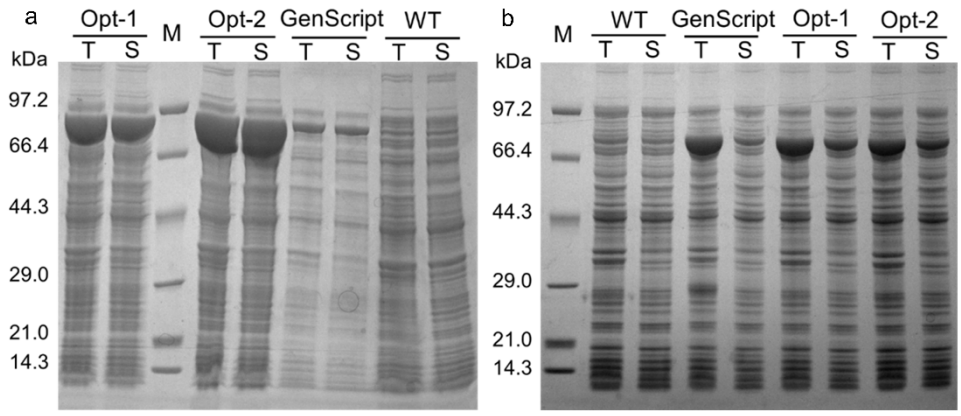

Figure 4. Representative SDS-PAGE analysis of recombinant proteins. Lane M, protein molecular weight marker; Opt, DNA sequences optimized using SCG-GELP; GenScript, GenScript-optimized DNA sequences; WT, wild-type sequences. T, whole-cell lysate; S, soluble fraction after centrifugation. (a) αGP (96.139 kDa). (b) IA (83.054 kDa). The complete DNA sequences of all tested variants are provided in the Supplementary Materials.

Table 1. Quantification of αGP and IA expression levels.

Protein	Protein	Solution type	Volumetric (Int)	Strip ratio (%)	Calculate
αGP	WT	T	317,115	5.6	-
		S	267,873	4.4	-
	GenScript	T	856,254	19.0	2.70
		S	810,231	27.0	3.02
	Opt-1	T	2,278,269	27.3	7.18
		S	2,005,263	25.0	7.49
	Opt-2	T	4,929,681	49.8	15.55

 doi:10.20944/preprints202605.1014.v1

IA	WT	S	5,496,312	47.7	20.52
		T	518,042	4.0	-
		S	622,202	4.8	-
	GenScript	T	4,288,970	31.8	8.28
		S	812,714	9.0	1.30
	Opt-1	T	3,480,176	26.6	6.72
		S	1,897,182	17.5	3.05
	Opt-2	T	3,115,910	22.2	6.01
		S	1,683,836	15.0	2.71

Note: The Calculate column represents the ratio of protein volume between the current protein and the corresponding WT solution type. Solution type T is the whole-cell lysate, and S is the solution in the supernatant of the centrifuged whole-cell lysate. Volumetric analysis and strip ratio are two parameters associated with protein gel analysis. The former involves the volumetric assessment of protein bands or spots, calculating the total signal quantity by integrating the pixel intensity of the bands, which reflects the total protein content within the band. The latter refers to the comparison of signal intensity ratios between different bands. The calculation is relative to WT for determining the volumetric fold change. Data shown are derived from a single proof-of-concept validation experiment.

2.5. DNA Sequence Analysis

To elucidate the molecular mechanisms by which SCG-GELP optimized sequences enhance soluble expression, we performed a codon-by-codon comparison of the wild-type (WT), GenScript-optimized, and SCG-GELP-optimized sequences for both αGP and IA. By comparing the codon compositions of the three sequences, we identified SCG-GELP-specific 'unique codon' sites that differed from both the WT and GenScript designs. These unique codons represent the key distinction between our strategy and existing commercial optimization approaches.

To characterize the distribution pattern of codon changes in SCG-GELP-optimized sequences, we tabulated two classes of "unique codon" sites (Figure 5a; see Supplementary Data File 1 for the complete list of unique codon sites). A unique codon was defined as a site simultaneously satisfying three criteria: (1) the codon differs from the wild-type (WT); (2) it differs from the GenScript (GS)-optimized sequence; and (3) both independent SCG-GELP-optimized sequences share the identical codon at this site. These sites were further subdivided based on whether GenScript altered the WT codon. Blue bars denote sites where GenScript retained the WT codon (GS = WT) but SCG-GELP introduced a redesigned codon; orange bars denote sites where both GenScript and SCG-GELP altered the WT codon but employed different substitutions. For αGP, the two categories comprised 80 and 107 sites, respectively; for IA, they comprised 125 and 94 sites, respectively.

Figure 5b illustrates the spatial distribution of amino acid residues corresponding to these unique codons (blue) within the three-dimensional protein structures. The top row shows two orientations of αGP, and the bottom row shows two orientations of IA. These residues were markedly enriched on the protein surface, whereas residues located in the hydrophobic core or within the active pocket were rarely affected, supporting a surface-targeted codon optimization strategy by SCG-GELP.

This preferential location of residues on the protein surface provides an important structural explanation for the experimental results. Traditional global codon optimization strategies, such as GenScript, typically maximize the Codon Adaptation Index (CAI) by distributing codon substitutions uniformly throughout the protein. While this approach can increase total expression, it fails to specifically improve folding efficiency, often leading to higher inclusion body formation. In contrast, the surface-targeted codon adjustments made by SCG-GELP likely improve soluble yields by coordinating translation elongation rates between surface and core regions, thereby optimizing co-translational folding [34]. This is clearly demonstrated in the IA validation: the GenScript-optimized

 doi:10.20944/preprints202605.1014.v1

sequence achieved only 9.0% soluble fraction, whereas IA-Opt1 reached 17.5%, nearly doubling the functional yield.

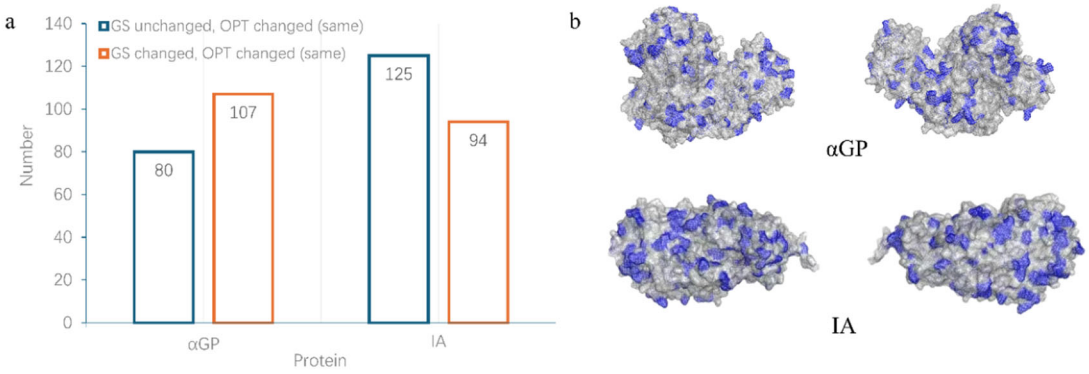

Figure 5. Statistical analysis of unique codon sites and their spatial distribution in protein structures. (a) Counts of two classes of unique codon sites. GS unchanged, OPT changed (same): GenScript retained the wild-type (WT) codon, whereas SCG-GELP introduced a different codon, with both optimized sequences identical at this site; GS changed, OPT changed (same): both GenScript and SCG-GELP altered the WT codon but employed different substitutions, with both optimized sequences matching each other. (b) Three-dimensional protein structures highlighting amino acid residues corresponding to unique codons (blue). Top row, αGP in two orientations; bottom row, IA in two orientations.

3. Discussion

The SCG-GELP framework offers an improved approach to codon optimization for recombinant protein expression in *E. coli*. By integrating a Transformer-based synonymous codon generation model with a multi-algorithmic expression prediction system, we have developed a solution that addresses key limitations of existing approaches. Experimental validation with αGP and IA demonstrates the framework's ability to substantially improve both total protein expression and soluble protein yield, outperforming commercial optimization tools while providing biologically interpretable sequence features.

The substantial improvements in soluble expression observed for both αGP and IA can be attributed to the synergy between large-scale sequence generation and expression-aware screening. While current transformer-based architectures generally perform well in codon optimization, the key advantage of SCG-GELP lies in its ability to generate a massive candidate pool and subsequently prioritize sequences with high soluble expression potential using experimentally validated data. This dual-step strategy effectively couples codon usage compliance with folding-favorable sequence features, enabling the discovery of surface-targeted optimization patterns that are inaccessible to conventional global optimization methods.

A key insight from this study is that SCG-GELP implicitly learns and executes a protein surface-targeted codon optimization strategy. To our knowledge, this represents the first computationally designed and experimentally validated framework for codon optimization specifically targeting surface amino acids in industrial enzyme engineering, offering new insights into the complex relationship between DNA sequence and soluble protein expression.

The framework's success can be attributed to several key factors. First, the use of large-scale, high-quality training data (98,855 *E. coli* sequences for SCG and 2,384 expression-annotated sequences for GELP) provides a robust foundation for learning. Second, the integration of multiple algorithmic approaches captures complementary aspects of sequence-expression relationships. Third, the rigorous experimental validation pipeline ensures practical relevance and reliability. These

 doi:10.20944/preprints202605.1014.v1

advantages position SCG-GELP as a valuable tool for both basic research and industrial enzyme production.

Several limitations and future directions warrant discussion. Expanding the training data to include more expression-annotated sequences from diverse hosts would enhance the model's versatility. While effective for single-gene optimization, SCG-GELP currently does not address multi-gene expression balancing—a critical requirement for pathway engineering. Additionally, the NESG dataset's semi-quantitative expression scores (0–5 scale) may lack the precision of quantitative proteomics, potentially limiting prediction granularity. Although SCG-GELP has demonstrated efficacy in *E. coli*, its generalizability to other expression hosts such as yeast or Bacillus subtilis requires further investigation. Additionally, the experimental validation was conducted as a single-batch proof-of-concept study. While the substantial improvements in soluble expression (>20-fold for αGP and nearly doubling the soluble fraction for IA) strongly support the efficacy of the framework, future work will include biological replicates to confirm the quantitative reproducibility of these results across independent experiments. Furthermore, while this study experimentally validated the SCG-GELP framework against wild-type sequences and a commercial optimization service (GenScript), a systematic benchmark comparison against other recently developed deep learning-based codon optimization tools was not performed. This is partly because many of these methods lack publicly available pre-trained models or standardized inference pipelines, making fair and reproducible comparisons technically challenging. We contend that for industrial enzyme engineering, the ultimate criterion for codon optimization is the actual soluble protein yield in vivo rather than computational metrics such as CAI or perplexity. In this regard, the experimental results presented here—particularly the substantial improvements over a commercially optimized sequence—provide a biologically meaningful assessment of the framework's practical utility. Nevertheless, comprehensive benchmarking on standardized datasets remains an important direction for future work.

From an industrial perspective, SCG-GELP offers tangible benefits for enzyme production. Its ability to substantially improve soluble expression—as demonstrated by the 20.52-fold increase for αGP—directly translates into reduced production costs. From a practical standpoint, the framework's computational efficiency enables rapid sequence optimization. The web interface (https://scg-gelp.biodesign.ac.cn) facilitates accessibility for researchers, while open-source availability of the code (https://github.com/yuddecho/SCG-GELP) promotes community adoption and further development.

4. Materials and Methods

4.1. Dataset Preparation

4.1.1. E. coli Genomic Dataset RefSeqE

All publicly available protein sequences and corresponding coding DNA sequences of *E. coli* were retrieved from the NCBI Reference Sequence (RefSeq) database and subjected to a comprehensive data cleaning process. The cleaning protocol comprised two principal phases: (1) elimination of sequences exhibiting inconsistencies between protein and DNA sequences based on the standard codon translation table; and (2) implementation of stringent quality criteria, requiring protein sequences to begin with a methionine residue (M), exclude ambiguous residues (X), end with a canonical stop codon (*), and lack internal stop codons (to avoid truncated proteins). Additionally, redundancy removal was performed using MMseqs2 [35,36] with a 30% sequence identity threshold (see Supplementary Methods for detailed commands and parameters). Given the large initial dataset (145 million sequences), batch-wise clustering was followed by merging and deduplication to obtain the final dataset, RefSeqE. We randomly selected 10,000 sequences from RefSeqE as the test set, with the remaining sequences reserved for training.

 doi:10.20944/preprints202605.1014.v1

Table 2. Changes in the number of sequences in the construction of the RefSeqE dataset.

Step	Number of sequences
Retrieval of RefSeq data from NCBI	144,775,103
Data cleaning	136,321,237
MMseqs2 30% clustering (RefSeqE dataset)	98,855
Train set: Test set	88,855: 10,000

Sequences were encoded using two distinct dictionaries for proteins and DNA. Each dictionary included four reserved tokens (0–3) representing <unk> (unknown character), <pad> (padding character), <bos> (beginning-of-sequence marker), and <eos> (end-of-sequence marker), respectively. The protein dictionary assigned numerical identifiers from 4 to 24 to the 20 canonical amino acids and the stop codon (*), with values inversely correlated to amino acid frequency (lower values indicate higher frequency; see Supplementary Table S1). Similarly, the DNA dictionary assigned identifiers from 4 to 67 to all 64 possible codons, with values determined by their frequencies in the dataset (see Supplementary Table S2).

4.1.2. Gene Expression Level Dataset NESG-DNA

Gene expression data were obtained from a large-scale high-throughput protein expression study [37–39], encompassing 6,348 genes from 2 eukaryotic species, 18 archaeal species, and 151 bacterial species. Proteins were systematically expressed and purified by the Northeast Structural Genomics (NESG) consortium. Whole-cell lysates and supernatants were subjected to SDS-PAGE analysis followed by Coomassie Brilliant Blue staining. Expression levels in whole-cell lysate (E) and supernatant (S) were quantified via visual inspection and assigned integer scores ranging from 0 (no expression) to 5 (maximal expression). Sequences with E and S scores ≤ 2 were classified as low-expression sequences, whereas sequences with E and S scores ≥ 4 were classified as high-expression sequences. As shown in Figure 6, a total of 1,313 low-expression (negative) and 1,071 high-expression (positive) DNA sequences were obtained to construct the NESG-DNA dataset.

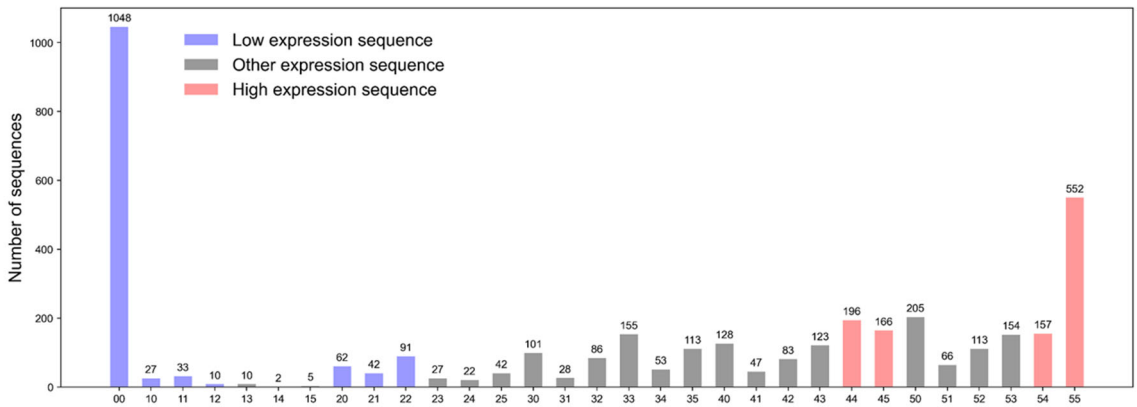

Figure 6. Distribution of gene expression level data. The horizontal coordinate markers indicate a combination of whole-cell fluid and supernatant expression levels, e.g., "54" indicates that the whole-cell fluid protein expression level (E) score is 5 and the supernatant protein expression level (S) score is 4.

4.2. Synonymous Codon Generation (SCG) Model

The synonymous codon generation (SCG) model adopts a classic encoder-decoder sequence-to-sequence (Seq2Seq) architecture inspired by natural language processing. The model comprises three core components. The first is an embedding layer that transforms sequence tokens into dense vector representations. The second comprises encoder and decoder modules: the encoder analyzes the input

 doi:10.20944/preprints202605.1014.v1

protein sequence to extract contextual features, producing a context vector that encapsulates the semantic information of the sequence; the decoder autoregressively generates output sequences based on this context vector and previously generated tokens. The third component is a linear projection layer that transforms decoder outputs into logits (unnormalized probabilities) over the codon vocabulary at each position. Both the encoder and decoder are implemented as Transformer networks [40] with positional encoding. The model hyperparameters are as follows: embedding dimensions of 512 for both encoder (vocabulary size 68) and decoder (vocabulary size 25); 3 Transformer layers; 8 attention heads; hidden layer dimension of 512; Adam optimizer with a learning rate of 0.001; and KLDivLoss (ignoring the <pad> index) as the training objective.

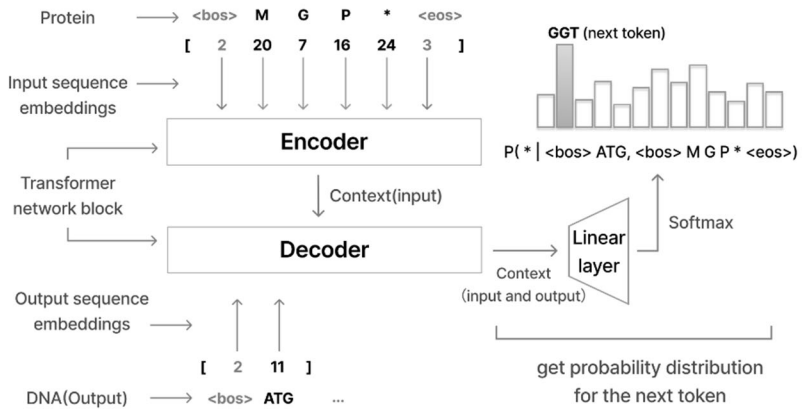

Figure 7. Process of synonymous codon generation using the SCG model. The protein sequence is first transformed via an embedding layer and serves as input to the encoder. The decoder receives both the contextual vectors generated by the encoder and the embedded representations of the partially generated DNA sequence. Through a linear layer followed by a Softmax layer, the model predicts the probability distribution over the next token, thereby generating the output sequence step by step.

The SCG model utilizes beam search [41] coupled with pruning operations during decoding. At each generation step, the model retains the top beam_size candidates ranked by unnormalized probabilities, while immediately discarding codons that would encode mismatched amino acids relative to the target protein sequence. The beam_size hyperparameter, which is bounded by the maximum number of synonymous codons for any amino acid (six for leucine), is typically set to 2 or 4 to balance search breadth and computational efficiency. Retained candidates are extended with new codons, and their cumulative log-probabilities are computed. These sequences populate a candidate pool with a predefined capacity (candidate_pool_size), which defaults to 1260. The pool is sorted by descending probability, and only the top candidate_pool_size entries are retained. Upon decoding completion, the model outputs these top-ranked sequences. Both parameters are user-configurable, ensuring flexibility in optimization granularity and output diversity.

4.3. Gene Expression Level Prediction (GELP) Model

The gene expression level prediction (GELP) model comprises two processing steps. First, DNA sequence features are extracted using the DNABERT-2 model. These features are then fed into an ensemble of classifiers—SVM, LR, and MLP—to predict the probability that a given sequence exhibits high soluble expression. The final output is the mean positive-class probability across the three classifiers.

The three classification algorithms, SVM, LR and MLP, were selected from 14 classic machine learning classification algorithms based on the NESG-DNA dataset using ten-fold cross validation (see Supplementary Figure S1 for the performance comparison of all 14 algorithms). We fine-tuned the DNABERT-2 model on a gene expression dataset (NESG-DNA) using default training

 doi:10.20944/preprints202605.1014.v1

parameters. The DNABERT-2 contains a Transformer language model consisting of 12 BertEncoder attention layers, with a hidden layer and output size of 768, based on upstream and downstream nucleotide contexts capturing a global and transferable understanding of genomic DNA sequences, and trained on large-scale multi-species genomes.

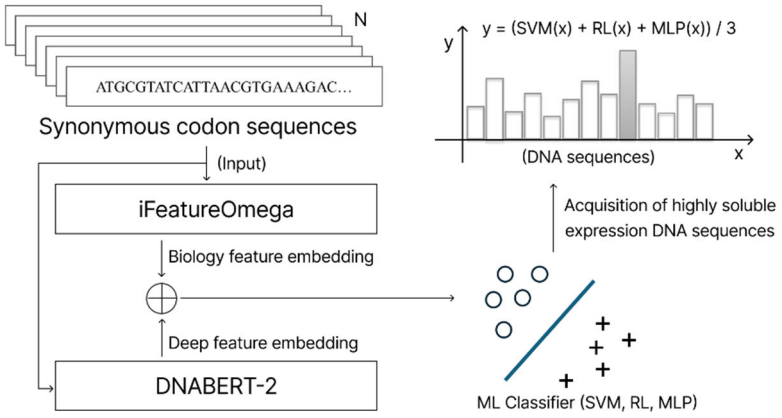

Figure 8. Screening process using the GELP model. The GELP model integrates biological and deep-learning features. The final prediction for identifying DNA sequences with high soluble expression is derived by averaging the independent predictions from a classifier ensemble (SVM, LR, MLP).

4.4. Experimental Validation

A proof-of-concept experimental validation was conducted to assess the expression performance of optimized sequences in a single batch. Each DNA sequence was cloned into the pET28a vector and transformed into *E. coli* BL21 (DE3) cells for protein expression. Single colonies grown on Terrific Broth (TB) agar plates were inoculated into 5 mL of TB liquid medium supplemented with 50 μg mL−1 kanamycin. The cultures were incubated at 37 °C with shaking until the optical density at 600 nm (OD600) reached 0.6–0.8. Protein expression was induced by adding isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 0.1 mM, followed by incubation at 16 °C for 20–22 h with continuous shaking. Cells were harvested by centrifugation at 6,000 × g for 5 min and resuspended in buffer A (50 mM HEPES, 50 mM NaCl, pH 7.5). Cell lysis was performed by sonication, and the lysate was clarified by centrifugation at 8,000 × g for 20 min. Protein expression was analyzed by 12% SDS-PAGE for both whole-cell lysates and soluble fractions. Gel images were captured using a Gel Doc XR+ imaging system, and protein expression levels were quantified using ImageLab software.

5. Conclusions

In conclusion, the SCG-GELP framework represents a notable step forward in computational protein expression optimization. By combining large-scale sequence learning with multi-algorithmic expression prediction and rigorous experimental validation, we have developed a tool that not only outperforms commercial optimization standards and wild-type sequences but also provides mechanistic insights into the complex relationship between DNA sequence and protein expression. Future work will focus on expanding the framework's applicability to additional host systems and integrating advanced sequence-structure features to further enhance predictive performance.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Author Contributions: Conceptualization, C.Y.; methodology, D.Y. and N.G.; software, D.Y., H.C., and R.W.; validation, L.F., Y.M.Q., and H.C.; formal analysis, D.Y. and S.S.; investigation, L.F. and Y.M.Q.; data curation,

 doi:10.20944/preprints202605.1014.v1

D.Y., N.G. and S.S.; writing—original draft, D.Y. and N,G; writing—review and editing, C.Y.; visualization, D.Y. and S.S.; supervision, C.Y. and X.L.; project administration, C.Y.; funding acquisition, C.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the article and supplementary materials, including the complete DNA sequences of tested variants, unique codon site statistics, and 14-classifier performance data. The source code and trained models are available at https://github.com/yuddecho/SCG-GELP.

Acknowledgments: The authors thank all colleagues who contributed to this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Doble, M.V.; Obrecht, L.; Joosten, H.J.; Lee, M.; Rozeboom, H.J.; Branigan, E.; Naismith, J.H.; Janssen, D.B.; Jarvis, A.G.; Kamer, P.C.J. Engineering thermostability in artificial metalloenzymes to increase catalytic activity. *ACS Catalysis* **2021**, *11*, 3620–3627, doi:10.1021/acscatal.0c05413.
2. Bernal, C.; Rodríguez, K.; Martínez, R. Integrating enzyme immobilization and protein engineering: an alternative path for the development of novel and improved industrial biocatalysts. *Biotechnology Advances* **2018**, *36*, 1470–1480, doi:https://doi.org/10.1016/j.biotechadv.2018.06.002.
3. Galloway, C.A.; Sowden, M.P.; Smith, H.C. Increasing the yield of soluble recombinant protein expressed in *E. coli* by induction during late log phase. *BioTechniques* **2003**, *34*, 524–530, doi:10.2144/03343st04.
4. Bhatwa, A.; Wang, W.; Hassan, Y.I.; Abraham, N.; Li, X.-Z.; Zhou, T. Challenges Associated With the Formation of Recombinant Protein Inclusion Bodies in *Escherichia coli* and Strategies to Address Them for Industrial Applications. *Frontiers in Bioengineering and Biotechnology* **2021**, *9*, 1–18, doi:10.3389/fbioe.2021.630551.
5. Nguyen, J.T.; Fong, J.; Fong, D.; Fong, T.; Lucero, R.M.; Gallimore, J.M.; Burata, O.E.; Parungao, K.; Rascón, A.A. Soluble expression of recombinant midgut zymogen (native propeptide) proteases from the aedes aegypti mosquito utilizing *E. coli* as a host. *BMC Biochemistry* **2018**, *19*, 1–14.
6. De Marco, A. Protocol for preparing proteins with improved solubility by co-expressing with molecular chaperones in *Escherichia coli*. *Nature Protocols* **2007**, *2*, 2632–2639.
7. Smialowski, P.; Doose, G.; Torkler, P.; Kaufmann, S.; Frishman, D. PROSO II--a new method for protein solubility prediction. *FEBS Journal* **2012**, *279*, 2192–2200, doi:10.1111/j.1742-4658.2012.08603.x.
8. Kramer, R.M.; Shende, V.R.; Motl, N.; Pace, C.N.; Scholtz, J.M. Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophysical Journal* **2012**, *102*, 1907–1915.
9. Levinthal, C. Are there pathways for protein folding? *Journal de Chimie Physique* **1968**, *65*, 44–45.
10. Hoover, D.M.; Lubkowski, J. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Research* **2002**, *30*, e43–e43, doi:10.1093/nar/30.10.e43.
11. Grote, A.; Hiller, K.; Scheer, M.; Münch, R.; Nörtemann, B.; Hempel, D.C.; Jahn, D. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Research* **2005**, *33*, W526–W531, doi:10.1093/nar/gki376.
12. Puigbo, P.; Guzman, E.; Romeu, A.; Garcia-Vallve, S. OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Research* **2007**, *35*, W126–W131, doi:10.1093/nar/gkm219.
13. Gaspar, P.; Moura, G.; Santos, M.A.S.; Oliveira, J.L. mRNA secondary structure optimization using a correlated stem–loop prediction. *Nucleic Acids Research* **2013**, *41*, e73–e73, doi:10.1093/nar/gks1473.
14. Villalobos, A.; Ness, J.E.; Gustafsson, C.; Minshull, J.; Govindarajan, S. Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics* **2006**, *7*, 285, doi:10.1186/1471-2105-7-285.

 doi:10.20944/preprints202605.1014.v1

15. Chin, J.X.; Chung, B.K.-S.; Lee, D.-Y. Codon Optimization OnLine (COOL): a web-based multi-objective optimization platform for synthetic gene design. *Bioinformatics* **2014**, *30*, 2210–2212, doi:10.1093/bioinformatics/btu192.
16. Guimaraes, J.C.; Rocha, M.; Arkin, A.P.; Cambray, G. D-Tailor: automated analysis and design of DNA sequences. *Bioinformatics* **2014**, *30*, 1087–1094, doi:10.1093/bioinformatics/btt742.
17. Chen, J. Strategies for high-level expression of recombinant protein in *Escherichia coli*. *China Biotechnology* **2007**, *27*, 103–109.
18. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589, doi:10.1038/s41586-021-03819-2.
19. Shin, J.-E.; Riesselman, A.J.; Kollasch, A.W.; McMahon, C.; Simon, E.; Sander, C.; Manglik, A.; Kruse, A.C.; Marks, D.S. Protein design and variant prediction using autoregressive generative models. *Nature Communications* **2021**, *12*, 2403, doi:10.1038/s41467-021-22732-w.
20. Alley, E.C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G.M. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods* **2019**, *16*, 1315–1322, doi:10.1038/s41592-019-0598-1.
21. Fu, H.; Liang, Y.; Zhong, X.; Pan, Z.; Huang, L.; Zhang, H.; Xu, Y.; Zhou, W.; Liu, Z. Codon optimization with deep learning to enhance protein expression. *Scientific Reports* **2020**, *10*, 17617, doi:10.1038/s41598-020-74091-z.
22. Jain, R.; Jain, A.; Mauro, E.; Leshane, K.; Densmore, D. ICOR: improving codon optimization with recurrent neural networks. *BMC Bioinformatics* **2023**, *24*, doi:10.1186/s12859-023-05246-8.
23. Li, S.; Moayedpour, S.; Li, R.; Bailey, M.; Riahi, S.; Kogler-Anele, L.; Miladi, M.; Miner, J.; Pertuy, F.; Zheng, D.; et al. CodonBERT large language model for mRNA vaccines. *Genome Res* **2024**, *34*, 1027–1035, doi:10.1101/gr.278870.123.
24. Fallahpour, A.; Gureghian, V.; Filion, G.J.; Lindner, A.B.; Pandi, A. CodonTransformer: a multispecies codon optimizer using context-aware neural networks. *Nature Communications* **2025**, *16*, 3205, doi:10.1038/s41467-025-58588-7.
25. Constant, D.A.; Gutierrez, J.M.; Sastry, A.V.; Viazzo, R.; Smith, N.R.; Hossain, J.; Spencer, D.A.; Carter, H.; Ventura, A.B.; Louie, M.T.M.; et al. Deep learning-based codon optimization with large-scale synonymous variant datasets enables generalized tunable protein expression. *bioRxiv preprint* **2023**, doi:10.1101/2023.02.11.528149.
26. Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning* **1995**, *20*, 273–297, doi:10.1007/BF00994018.
27. Zhihan Zhou; Yanrong Ji; Weijian Li; Pratik Dutta; Ramana Davuluri; Liu, H. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. *arXiv preprint:2306.15006* **2023**.
28. Sharp, P.M.; Li, W.-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* **1987**, *15*, 1281–1295, doi:10.1093/nar/15.3.1281.
29. Kane, J.F. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Current Opinion in Biotechnology* **1995**, *6*, 494–500, doi:https://doi.org/10.1016/0958-1669(95)80082-4.
30. Song, Y.H.; Wu, R.R.; Wei, X.L.; Shi, T.; Li, Y.J.; You, C.; Zhang, L.L.; Zhu, Z.G.; Zhang, Y.H. Advances in a new energy system based on electricity-hydrogen-carbohydrate cycle. *Sheng Wu Gong Cheng Xue Bao* **2022**, *38*, 4081–4100.
31. Zhu, Z.G.; Kin Tam, T.; Sun, F.F.; You, C.; Zhang, Y.H. A high-energy-density sugar biobattery based on a synthetic enzymatic pathway. *Nature Communications* **2014**, *5*, 3026, doi:10.1038/ncomms4026.
32. You, C.; Shi, T.; Li, Y.J.; Han, P.P.; Zhou, X.G.; Zhang, Y.H. An in vitro synthetic biology platform for the industrial biomanufacturing of myo-inositol from starch. *Biotechnology and Bioengineering* **2017**, *114*, 1855–1864.
33. Cheng, K.; Zhang, F.; Sun, F.; Chen, H.; Percival Zhang, Y.H. Doubling Power Output of Starch Biobattery Treated by the Most Thermostable Isoamylase from an Archaeon *Sulfolobus tokodaii*. *Scientific Reports* **2015**, *5*, 13184, doi:10.1038/srep13184.

 doi:10.20944/preprints202605.1014.v1

34. Buhr, F.; Jha, S.; Thommen, M.; Mittelstaet, J.; Kutz, F.; Schwalbe, H.; Rodnina, M.V.; Komar, A.A. Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Molecular Cell* **2016**, *61*, 341–351, doi:10.1016/j.molcel.2016.01.008.
35. Mirdita, M.; Steinegger, M.; Söding, J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* **2019**, *35*, 2856–2858, doi:10.1093/bioinformatics/bty1057.
36. Steinegger, M.; Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **2017**, *35*, 1026–1028, doi:10.1038/nbt.3988.
37. Boël, G.; Letso, R.; Neely, H.; Price, W.N.; Wong, K.-H.; Su, M.; Luff, J.D.; Valecha, M.; Everett, J.K.; Acton, T.B. Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature* **2016**, *529*, 358–363.
38. Price, W.N.; Handelman, S.K.; Everett, J.K.; Tong, S.N.; Bracic, A.; Luff, J.D.; Naumov, V.; Acton, T.; Manor, P.; Xiao, R. Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in *E. coli*. *Microbial Informatics and Experimentation* **2011**, *1*, 1–20.
39. Hon, J.; Marusiak, M.; Martinek, T.; Kunka, A.; Zendulka, J.; Bednar, D.; Damborsky, J. SoluProt: prediction of soluble protein expression in *Escherichia coli*. *Bioinformatics* **2021**, *37*, 23–28, doi:10.1093/bioinformatics/btaa1102.
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv preprint arXiv:1706.03762* **2017**.
41. Meister, C.; Vieira, T.; Cotterell, R. Best-First Beam Search. *arXiv preprint arXiv:2007.03909* **2020**.