Supplement for NatCodon: protein-structure-informed *de novo* codon sequence design for efficient protein expression

Yujie Qin[1,2,*], Kaiyue Wang[1,2,*] , Mengchen Pu[3], Zilin Lyu[3], Zibo Cui[1,2], Fenghui Shi[1,2], & Shen Ge[3,#], Sheng Ye[1,2,4,#]

[1] School of Biomedical Science and Medicine Engineering, Beihang University, Beijing 100191, People's Republic of China

[2] Key Laboratory of Precision Diagnosis and Treatment Technology and Intelligent Imaging Equipment (Beihang University), Ministry of Industry and Information Technology, Beijing 100862, China

[3] dProtein Biotechnologies Co. Ltd., Beijing 102206, China

[4] School of Engineering Medicine, Beihang University, Beijing 100191, China

* These authors contributed equally to the work.

[#] Correspondence should be addressed to: yesheng@buaa.edu.cn (S. Ye), sge@dprotein.cn

Contents

I Supplementary Information Figures 4
SI Figure 1 5
SI Figure 2 7
SI Figure 3 9
SI Figure 4 11
SI Figure 5 12
SI Figure 6 13
II Supplementary Methods 14
1. NatCodon: model architecture and training details 15
1.1. Data curation 15
1.2. Model architecture 16
1.3. Training setup 17
1.4. Loss function 17
1.5. Codon Gate 19
1.6. Focus-controlled sampling strategy 20
2. Structure–codon coupling analysis 22
2.1. Mutual information 22
2.2. χ^2 tests for single-residue 23
2.3. Structure-specific codon enrichment analysis 25
2.4. tRNA abundance 25
3. In silico experimental validation 27
3.1. Similarty 27
3.2. Jaccard Index 27
3.3. CAI calculation 28
3.4. GC content 29
3.5. tAI calculation 29
3.6. %MinMax 30
3.7. Kinetic distances 31
3.8. Principal component analysis 32
3.9. Structure-conditioned codon divergence 33
3.10. NatCodon Idex 35
3.10.1. Structure-aware CAI 35
3.10.2. Single-chain SCCD 36
3.10.3. Final NCI formulation 36
4. Statistical analysis 38

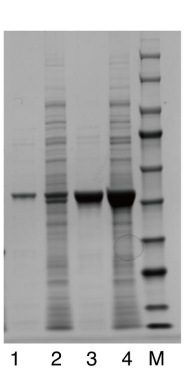
4.1. Cliff's δ 38
4.2. Wasserstein distance. 38
4.3. Spearman rank correlation 39
5. In vitro experimental validation 40
5.1. Selection of Experimental Subjects 40
5.2. Gene synthesis and cloning 47
5.3. Protein expression and solubility analysis 47
5.4. Protein purification 49
5.5. Mutagenesis experiments 49
6. Model Explanation 50
6.1. SHAP 50
6.2. IG attribution 50
6.3. Positional profiles of tAI 51
Reference 53

Part I

Supplementary Information Figures

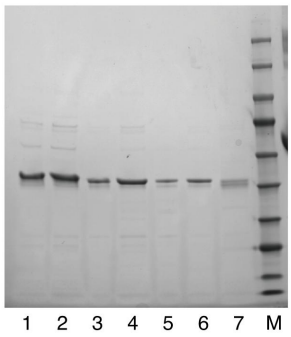
SI Figure 1

A



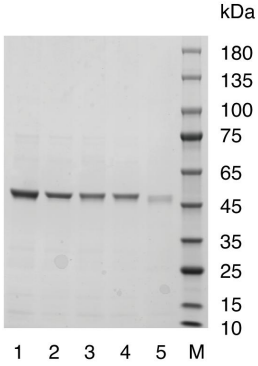
Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
1	DHOase-OP-250	46.36	0.01
2	DHOase-OP-250+20	46.36	0.04
3	DHOase-NC-1-250	46.36	0.33
4	DHOase-NC-1-250+20	46.36	0.46

B



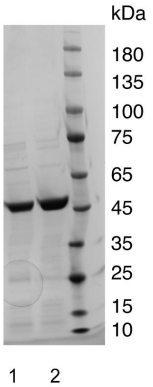
Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
1	DHOase-NC-2	46.36	0.13
2	DHOase-NC-3	46.36	0.17
3	DHOase-NC-4	46.36	0.01
4	DHOase-NC-1	46.36	0.13
5	DHOase-NC-NS2	46.36	0.00
6	DHOase-NC-NS1	46.36	0.02
7	DHOase-OP-1	46.36	0.01

C



Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
1	DHOase-NC-1	46.36	0.20
2	DHOase-NC-2	46.36	0.07
3	DHOase-NC-3	46.36	0.09
4	DHOase-NC-4	46.36	0.05
5	DHOase-OP-1	46.36	0.00

D

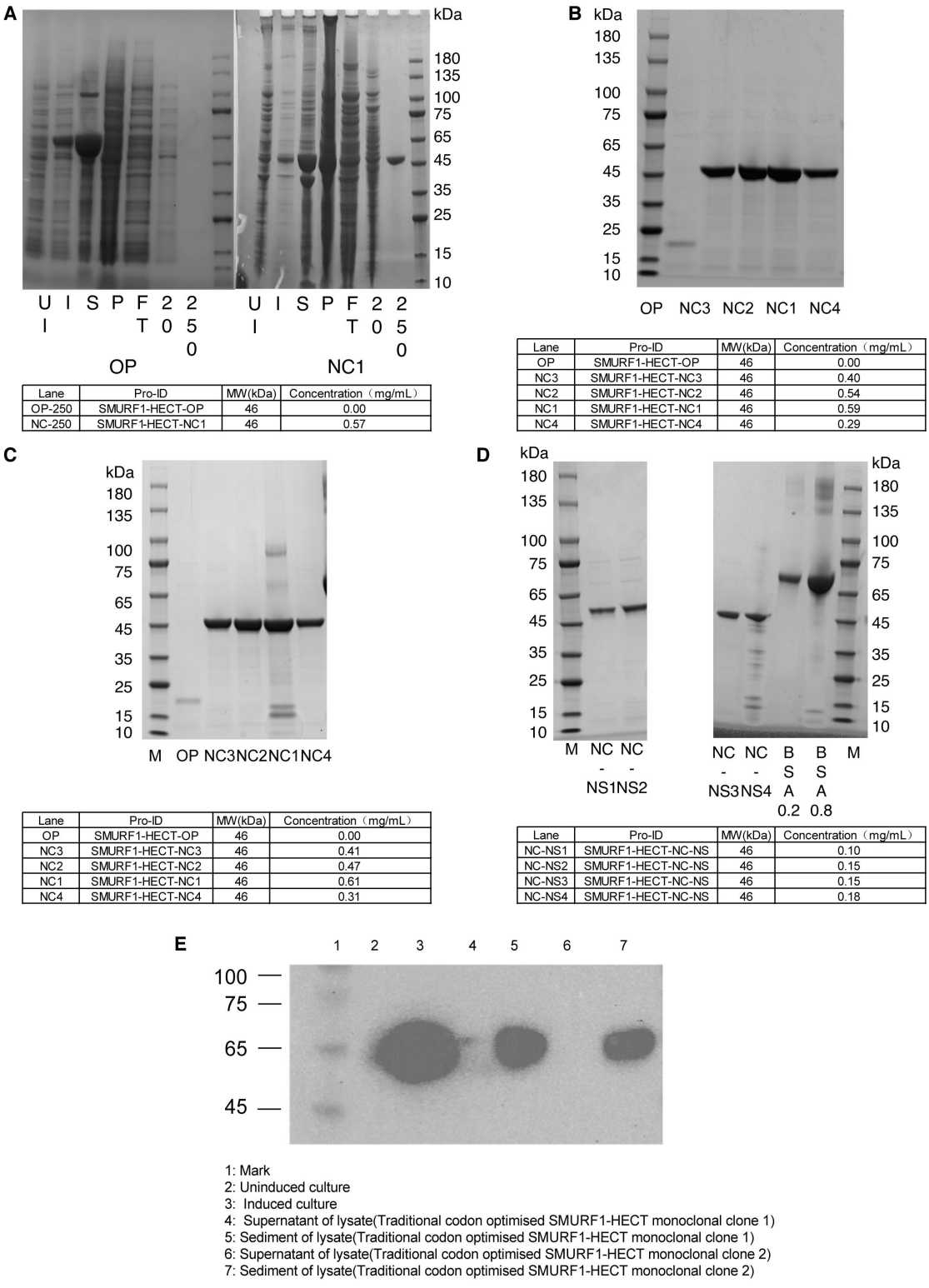


Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
1	DHOase-NC-NS	46.36	0.20
2	DHOase-NC-NS	46.36	0.25

SI Figure 1: DHOase. A. Preliminary screening of DHOase-OP and NC-1 variants. Lanes 1–4 correspond to DHOase-OP-250, DHOase-OP-250+20, DHOase-NC-1-250,

and DHOase-NC-1-250+20, respectively. B. Purity assessment of multiple DHOase variants (NC series and OP series). Lanes 1 – 7 show NC-2, NC-3, NC-4, NC-1, NC-NS2, NC-NS1, and OP-1, respectively. C. Batch-to-batch comparison of DHOase variants. Lanes 1 – 5 correspond to NC-1, NC-2, NC-3, NC-4, and OP-1 from an independent batch. NC-1 (0.20 mg/mL) shows the most intense band, whereas OP-1 (0.00 mg/mL) shows no detectable band, demonstrating good reproducibility. D. Validation of DHOase-NC-NS variants. Lanes 1 – 2 represent two independently prepared DHOase-NC-NS samples, with concentrations of 0.20 mg/mL and 0.25 mg/mL, respectively.

87 SI Figure 2



Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
OP-250	SMURF1-HECT-OP	46	0.00
NC-250	SMURF1-HECT-NC1	46	0.57

Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
OP	SMURF1-HECT-OP	46	0.00
NC3	SMURF1-HECT-NC3	46	0.40
NC2	SMURF1-HECT-NC2	46	0.54
NC1	SMURF1-HECT-NC1	46	0.59
NC4	SMURF1-HECT-NC4	46	0.29

Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
OP	SMURF1-HECT-OP	46	0.00
NC3	SMURF1-HECT-NC3	46	0.41
NC2	SMURF1-HECT-NC2	46	0.47
NC1	SMURF1-HECT-NC1	46	0.61
NC4	SMURF1-HECT-NC4	46	0.31

Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
NC-NS1	SMURF1-HECT-NC-NS	46	0.10
NC-NS2	SMURF1-HECT-NC-NS	46	0.15
NC-NS3	SMURF1-HECT-NC-NS	46	0.15
NC-NS4	SMURF1-HECT-NC-NS	46	0.18

1: Mark
2: Uninduced culture
3: Induced culture
4: Supernatant of lysate(Traditional codon optimised SMURF1-HECT monoclonal clone 1)
5: Sediment of lysate(Traditional codon optimised SMURF1-HECT monoclonal clone 1)
6: Supernatant of lysate(Traditional codon optimised SMURF1-HECT monoclonal clone 2)
7: Sediment of lysate(Traditional codon optimised SMURF1-HECT monoclonal clone 2)

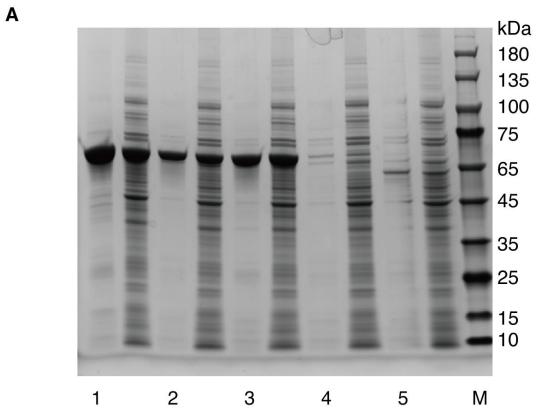
88

89 SI Figure 2: $SMURF1_{HECT}$. A. Induced expression and solubility analysis of

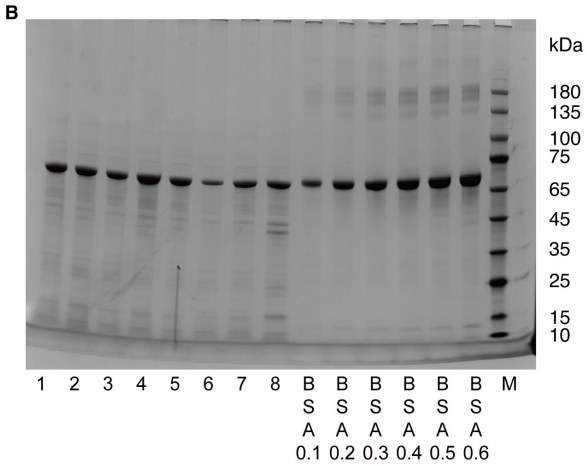
90 $SMURF1_{HECT}$ variants (OP and NC1). Expression levels of sequences optimized by

conventional methods (OP) and the new strategy (NC1) are compared. Lane labels: UI (uninduced whole cells), I (induced whole cells), S (supernatant/soluble fraction), P (pellet/inclusion bodies), FT (flow-through), and 20 and 250 (elution fractions with 20 mM and 250 mM imidazole, respectively). B-C. Parallel comparison of purification profiles for different NC variants (NC1–NC4) and the OP group. Results from two independent batches (B and C) are shown. The theoretical molecular weight of $SMURF1_{HECT}$ variants is ~46 kDa. D. Purification analysis of $SMURF1_{HECT}$ derived variants (NC-NS1 to NC-NS4). The purification yields of four non-synonymous variants are presented. E. Induced expression and fractionation analysis of the $SMURF1_{HECT}$ protein.

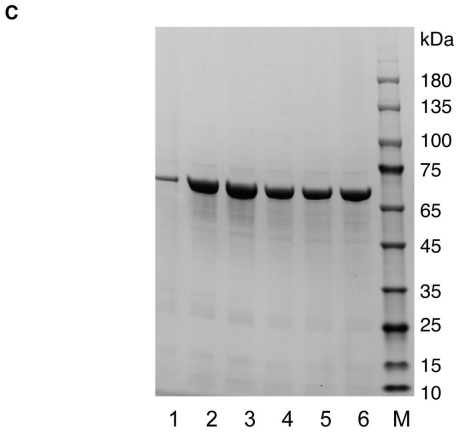
SI Figure 3



Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
1	Luci-NC1	61.50	0.23
2	Luci-OP	61.50	0.06
3	Luci-NC2	61.50	0.12
4	Luci-NC-NS	61.50	0.00
5	Luci-NC-NS	61.50	0.00



Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
1	Luci-NC5	61.50	0.20
2	Luci-NC7	61.50	0.18
3	Luci-NC2	61.50	0.14
4	Luci-NC4	61.50	0.26
5	Luci-NC6	61.50	0.15
6	Luci-OP	61.50	0.03
7	Luci-NC1	61.50	0.13
8	Luci-NC3	61.50	0.17

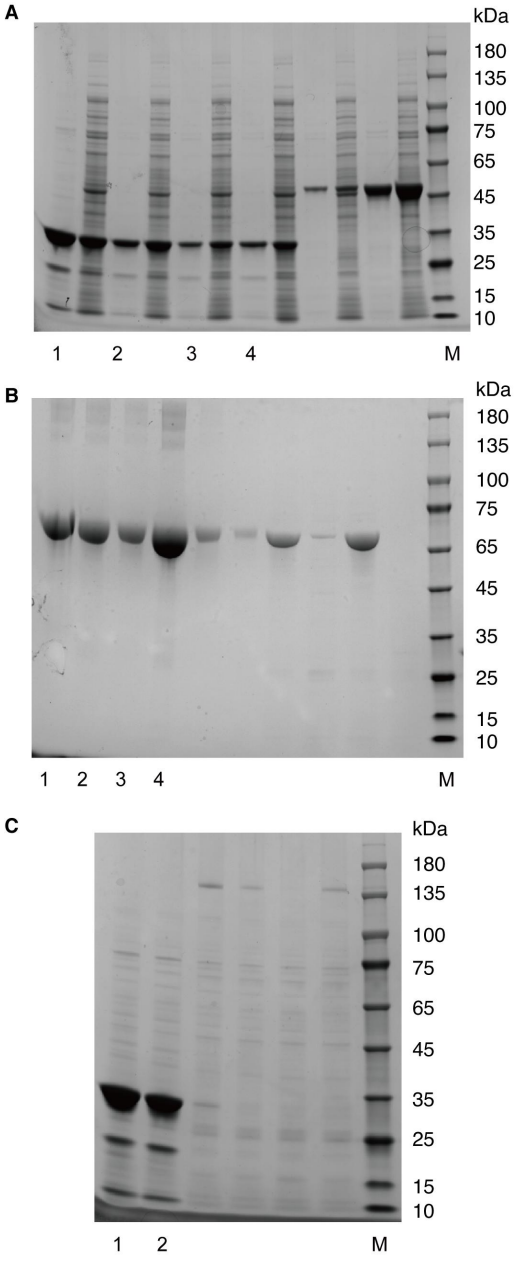


Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
1	Luci-OP	61.50	0.01
2	Luci-NC1	61.50	0.24
3	Luci-NC7	61.50	0.29
4	Luci-NC6	61.50	0.16
5	Luci-NC5	61.50	0.14
6	Luci-NC3	61.50	0.19

SI Figure 3: Luciferase. A. Preliminary screening of luciferase (Luci) variant expression. NC1, NC2, OP, and NC-NS variants are compared. B. Parallel comparison and semi-quantitative analysis of multiple Luci-NC variants. Lanes 1–8 correspond to NC1–NC7 and the OP variant. C. Batch validation of Luci variant

purification. Purification performance of OP and the NC variants is shown.

SI Figure 4



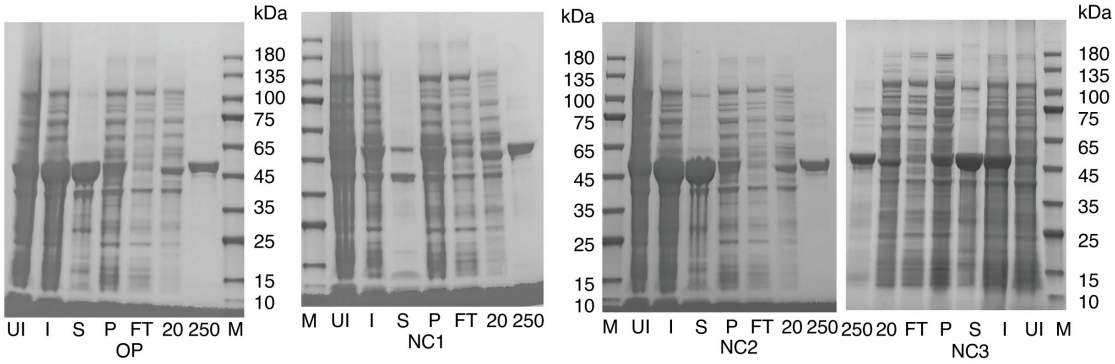
Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
1	GFP-NC1	27.7	0.48
2	GFP-NC2	27.7	0.18
3	GFP-NC3	27.7	0.07
4	GFP-OP	27.7	0.13

Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
1	GFP-OP	27.7	0.26
2	GFP-NC2	27.7	0.25
3	GFP-NC3	27.7	0.13
4	GFP-NC1	27.7	0.63

Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
1	GFP-NC1	27.7	0.57
2	GFP-OP	27.7	0.41

SI Figure 4: GFP. A. Preliminary screening of GFP variant expression. Lanes 1–4 correspond to NC1, NC2, NC3, and OP, respectively. B. Comparison of purified products from multiple GFP variants. Lanes 1–4 correspond to OP, NC2, NC3, and NC1, respectively. C. Direct comparison of production yield between NC1 and OP variants. Lane 1 corresponds to NC1, and lane 2 corresponds to OP.

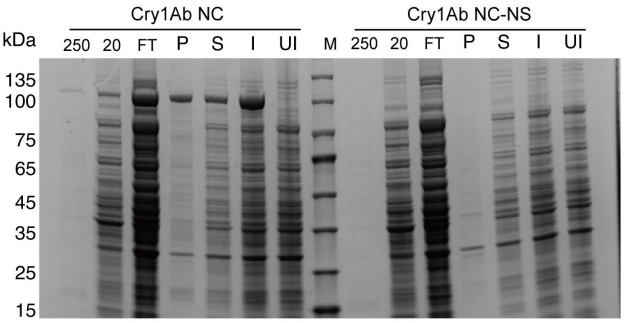
SI Figure 5



Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
1	HPPD-OP	44.7	0.56
3	HPPD-NC1	44.7	1.26
4	HPPD-NC2	44.7	0.87
2	HPPD-NC3	44.7	1.22

SI Figure 5: Effects of different codon optimization strategies on the expression level, solubility, and purification efficiency of the HPPD protein (theoretical molecular weight: 44.7 kDa). UI (Uninduced): uninduced whole-cell sample; I (Induced): whole-cell sample after IPTG induction; S (Soluble): supernatant after cell lysis (soluble protein fraction); P (Pellet): precipitate after cell lysis (insoluble/inclusion body fraction); FT (Flow-Through): flow-through fraction from Ni-affinity chromatography; 20: fraction obtained from washing with 20 mM imidazole buffer; 250: fraction eluted with 250 mM imidazole buffer containing the target protein.

SI Figure 6



Lane	Pro-ID	MW(kDa)	Concentration（mg/mL）
1	Cry1Ab-NC-NS	130.6	0.00026
2	Cry1Ab-NC1	130.6	0.00000

SI Figure 6: SDS-PAGE analysis of expression and purification of Cry1Ab variant proteins. UI (Uninduced): uninduced whole-cell sample; I (Induced): whole-cell sample after IPTG induction; S (Soluble): supernatant after cell lysis (soluble protein fraction); P (Pellet): precipitate after cell lysis (insoluble/inclusion body fraction); FT (Flow-Through): flow-through fraction from Ni-affinity chromatography; 20: fraction obtained from washing with 20 mM imidazole buffer; 250: fraction eluted with 250 mM imidazole buffer containing the target protein.

Part II

Supplementary Methods

1. NatCodon: model architecture and training details

1.1. Data curation

A non-redundant dataset of bacterial proteins was constructed by linking entries from the Protein Data Bank (PDB)[1] and NCBI GenBank[2]. We first selected all bacterial proteins with structures determined by X-ray crystallography to a resolution of 3.0 Å or better, yielding an initial pool of 60,306 entries. For each structure, the corresponding coding sequence was retrieved from the NCBI database using the '1D Coordinate Server' from RCSB PDB. A central hypothesis of this work is that the translational rhythm guiding CTF is encoded via structure-specific codon usage. Because secondary structures often form early during CTF, they provide the context in which to investigate this encoded signal[3,4]. Therefore, to test this hypothesis, we annotated each residue in our dataset with its structural context. Secondary structure assignments (e.g., H for α-helix, E for β-sheet) were generated from the PDB coordinates using the DSSP algorithm (Version 4.4)[5]. This annotation is fundamental to our model, as it allows us to move beyond global codon statistics and probe the localized, structure-dependent codon preferences that nature may have evolved to ensure folding fidelity. Finally, a rigorous filtering pipeline was applied to ensure data integrity. Sequences were first clustered at a 90% identity threshold using MMseqs2[6]

to create a non-redundant set, preventing biases from over-represented protein families. Subsequently, we discarded any entry where the amino acid sequence translated from the CDS did not perfectly match the sequence reported in the PDB file. This stringent curation process yielded a final high-quality dataset of 45,126 protein-structure-gene triplets for model training and evaluation.

1.2. Model architecture

NatCodon is a Transformer-based [7] generative model that learns the structure-codon grammar of CTF. A core innovation is its input, which consists of paired amino acid and secondary structure sequences. These paired tokens are first mapped into a shared high-dimensional embedding space via learned embedding layers. To encode positional information and enable effective modeling of long-range dependencies, we employ rotary positional embeddings (RoPE), which are applied directly within the self-attention mechanism to provide relative positional awareness without introducing additional positional parameters. The resulting sequence of embeddings is processed using a decoder-only Transformer architecture, following the design principles of LLaMA-style language models. Specifically, the model consists of a stack of four LLaMA decoders, each comprising multi-head self-attention and

feed forward SwiGLU layers. Through self-attention, the model captures complex dependencies across amino acid sequence, structural context, and positional information, allowing it to learn a conditional distribution over synonymous codons given both sequence and structure. To generate biologically valid and tunable codon sequences, the architecture incorporates three crucial components: a custom loss function, a "Codon Gate" and a focus-controlled sampling strategy.

1.3. Training setup

The model was trained for 100 epochs, with convergence typically achieved after approximately 20 epochs, as assessed by stabilization of the training loss. All experiments were conducted on an NVIDIA A100-SXM4-80GB GPU using CUDA 12.4.

1.4. Loss function

To adapt to the specific task of codon design, we have designed a custom Label Smooth Loss. The loss function consists of negative log-likelihood loss (L_{NLL}), codon loss(L_{codon}), and smoothing loss(L_{smooth}):

$$L_{Label\ smooth} = L_{NLL} + L_{codon} + L_{smooth}. \quad (1)$$

Specifically, let the batch size be denoted as B, the total number of codon categories be denoted as C, and the length of the sample be denoted as L.

L_{NLL}: The negative log-likelihood (NLL) loss quantifies the discrepancy between the predicted codon distribution and the ground-truth codon sequence. It is given by

$$\mathrm{L_{NLL}} = \frac{\sum_{b=1}^{B}\sum_{l=1}^{L}\log \mathrm{P}(y_{b,l}|x_{b.l})}{\mathrm{N_{entry}}}, \tag{2}$$

where $y_{b,l} \in \boldsymbol{y}$, represents the target label, $x_{b,l} \in \boldsymbol{x}$ represents the model output, $\mathrm{N_{entry}}$represents the total number of unmasked elements.

L_{codon}: A cross-entropy loss that introduces a frequency-based weighting scheme to correct codon imbalance, reduces bias toward overrepresented codons and promoting balanced learning across synonymous codon classes. It is given by

$$L_{codon} = \frac{\sum_{b=1}^{B}\sum_{c=1}^{C}\log \mathrm{P}(\mathrm{Codon_c}|x_b) \bullet \mathrm{w_{codon}}}{\mathrm{B} \bullet 64}, \tag{3}$$

where $\log \mathrm{P}(\mathrm{Codon_c}|x_b)$ represents the output of x_b, the predicted probability of the c-th codon class $\mathrm{w_{codon}}$ can be calculated by $\mathrm{w_{codon}} = \frac{y_{\mathrm{onehot}}}{\sum_{c=1}^{C} y_{\mathrm{onehot}}}$, which can dynamically adjust according to the distribution of the labels, helping to avoid overrepresentation of certain codon classes.

L_{smooth}: An entropy-inspired smoothing loss designed to encourage a smoother output distribution over codon classes, reducing overconfidence and improving generalization. It is given by

$$L_{smooth} = \frac{\sum_{b=1}^{B}\sum_{l=1}^{L}\log \mathrm{P}(\bullet|x_{b.l})}{N_{length}}, \tag{4}$$

where $\mathrm{P}(\cdot|x_{b.l})$ represents the predicted probability of each possible class given the

model output $x_{b,l}$, and N_{length} refers to the sequence length before padding. Label smooth loss can be adjusted based on the frequency or importance of different codons, and effectively controls the noise impact during model training through L_{smooth}, preventing overfitting and improving the model's generalization ability. This loss function, through weight adjustment and noise control, optimizes the quality and efficiency of protein codon design, making it especially suitable for cross-species expression optimization tasks.

1.5. Codon Gate

Despite the vast training data and the model's predictive power, the probabilistic nature of a standard Transformer output means that the generated codons could, on occasion, fail to translate to the correct amino acid. To eliminate this possibility and enforce absolute biological accuracy, we introduced the Codon Gate—a dynamic gating mechanism that ensures all generated sequences are valid. Positioned between the final Transformer layer and the output softmax function, the gate acts as a context-specific filter on the model's raw predictions. For each amino acid position in the input sequence, the Codon Gate first identifies the set of all valid, synonymous codons for that specific residue. It then constructs a dynamic mask that selectively "opens the gate" for the logits corresponding to these valid codons, while decisively

"closing the gate" for all remaining non-synonymous codons by setting their logit values to negative infinity. This operation effectively prunes the solution space in real-time, forcing the subsequent softmax function to distribute probability mass only among the biologically permissible options. This gating mechanism is critical: it guarantees that the final generated sequence is always a perfect, in-frame translation of the target protein, completely eliminating the possibility of non-synonymous errors or "hallucinations"[8]. By doing so, it decouples the model's complex task of learning contextual codon preferences from the simple, absolute requirement of biological accuracy, leading to a more stable and reliable design process.

1.6. Focus-controlled sampling strategy

Standard decoding via argmax deterministically selects the codon with the highest logit value. However, in cases where multiple synonymous codons have very similar high-ranking logits, this approach can be sensitive to minor model fluctuations and may discard near-optimal, potentially beneficial alternatives[9,10]. To address this and provide tunable control over sequence diversity, we developed a Focus-Controlled Sampling strategy. This strategy enables a dynamic balance between exploiting the model's highest-confidence predictions and exploring a constrained set of high-probability alternatives. Unlike global temperature scaling, this method operates

on a position-specific, relative-confidence basis. First, for each position, the algorithm identifies the single most likely codon and its corresponding probability P_{max}. A dynamic probability threshold is then calculated by scaling this maximum probability by a user-defined focus parameter, ϕ (ranging from 0 to 1). Subsequently, the candidate pool is pruned by setting the probability of any codon with a probability below this threshold (ϕ * P_{max}) to zero. A final codon is then drawn via multinomial sampling [11] from this refined, high-confidence distribution. A focus parameter of $\phi = 1.0$ results in deterministic, greedy selection of the top codon, while values closer to 0 allow for controlled, localized exploration of high-probability alternatives without venturing into low-confidence, high-risk codon choices.

2. Structure–codon coupling analysis

To systematically quantify the extent to which codon usage is modulated by protein structural context, we formulated codon selection as a conditional probabilistic process dependent on local structural states, and evaluated the statistical dependency between codon identity and structural annotation.

Let C denote the synonymous codon identity at a given residue position and S denote the corresponding structural state derived from protein structure annotation. Each codon sequence was thus represented as a sequence of paired observations $\{(c_i, s_i)\}_{i=1}^{L}$, where L is the protein length. The objective is to determine whether the distribution $P(C|S)$ deviates significantly from the marginal distribution $P(C)$, indicating that codon choice is conditioned on structural context rather than being independently sampled. To eliminate confounding effects from amino acid composition, each codon observation was associated with its corresponding amino acid identity A, and downstream analyses were performed either within amino acid–specific subsets or by explicitly conditioning on A.

2.1. Mutual information

We quantified the dependency between codon identity and structural state using mutual information (MI), defined as:

$$I(C;S) = \sum_{c \in C, s \in S} p(c,s) \log \frac{p(c,s)}{p(c)p(s)}. \quad (5)$$

Empirical joint and marginal distributions were estimated via frequency counts over all valid residue positions.

For each dataset, we generated 1,000 permuted samples and recomputed $I(C;S)$ to obtain a null distribution. Statistical significance was assessed by computing empirical p-values as:

$$p = \frac{(MI_{perm} \geq MI_{real}) + 1}{N_{perm} + 1}, \quad (6)$$

where MI_{real} is the observed mutual information and MI_{perm} are values from permuted datasets, N_{perm} is the number of permuted samples. This conservative formulation avoids zero p−values and ensures robustness.

2.2. χ^2 tests for single-residue

To quantify the coupling between protein structure and synonymous codon usage at residue-level, we performed the χ^2 test to each amino acid-specific synonymous codons to test the null hypothesis that codon usage is independent of protein structure. For each amino acid A, we constructed a contingency table summarizing the joint distribution of codon identity and structural state, where rows correspond to structural categories and columns correspond to synonymous codons. Each entry in the table

represents the observed count of a codon occurring within a specific structural context.

We then applied the χ^2 test to each amino acid–specific contingency table to assess whether the observed joint distribution deviates from that expected under independence. The test statistic is defined as:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{7}$$

where O_{ij} and E_{ij} denote the observed and expected counts for structure i and codon j, respectively, and expected counts are computed assuming independence between C and S given S.

To ensure the validity of the asymptotic approximation, we quantified the fraction of cells with expected counts below 5 for each contingency table and report this as a diagnostic measure. Amino acids with fewer than two synonymous codons were excluded from statistical testing.

Because the χ^2 statistic is sensitive to sample size, we further quantified the strength of association using Cramér's V, defined as:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}, \tag{8}$$

where n is the total number of observations for a given amino acid and k is the smaller dimension of the contingency table. This normalized metric provides an effect

size that is comparable across amino acids with different sample sizes and codon degeneracies. Together, these analyses enable the detection (via χ^2) and quantification (via Cramér's V) of structure-dependent biases in codon usage at single-residue resolution.

2.3. Structure-specific codon enrichment analysis

To further characterize the nature of structure-dependent codon usage, we computed enrichment scores for each codon–structure pair using log-odds ratios:

$$log\frac{P(c|s)}{P(c)}. \quad (9)$$

Positive values indicate codon enrichment in a given structural state, whereas negative values indicate depletion. These values were visualized as heatmaps to reveal systematic patterns of codon preference across structural categories.

2.4. tRNA abundance

To quantify the contribution of tRNA availability to codon usage bias, we estimated the relative adaptiveness of each codon using a modified tRNA adaptation index (tAI) framework. Anticodon copy numbers were obtained from the Genomic tRNA Database (https://gtrnadb.ucsc.edu) for *E. coli*. These copy numbers serve as a proxy for intracellular tRNA abundance, under the assumption that gene copy number correlates with tRNA concentration.

Each sense codon c was associated with one or more cognate or wobble-recognizing anticodons. Codon–anticodon pairing rules were defined based on canonical Watson–Crick pairing and wobble base pairing constraints. For each codon–anticodon pair, a selective constraint coefficient s_{ij} was assigned to reflect the relative efficiency of decoding, where perfect Watson–Crick pairing was assigned higher weights and wobble interactions were downweighted.

For each codon c, the absolute adaptiveness w_i was calculated as:

$$w_i = \sum_j s_{ij} \bullet n_j, \tag{10}$$

where n_j denotes the gene copy number of anticodon j, and s_{ij} represents the pairing efficiency between codon iii and anticodon j. This formulation integrates both tRNA abundance and decoding efficiency into a single measure of codon translational adaptiveness.

To normalize codon adaptiveness within each amino acid family and remove biases due to amino acid composition, we computed the relative adaptiveness W_i as:

$$W_i = \frac{w_i}{\max\limits_{c' \in C_A} w_{c'}}, \tag{11}$$

where C_A denotes the set of synonymous codons encoding amino acid A. This normalization constrains W_i to the range [0,1], where 1 corresponds to the most adapted codon for a given amino acid.

3. In silico experimental validation

3.1. Similarty

To quantify the similarity between generated codon sequences and their corresponding natural counterparts, we computed a codon-level sequence identity metric. For each protein, the CDSs from the natural reference and model-generated codon sequences were first segmented into non-overlapping codons of length three. Let $C^{(nat)} = (c_1^{(nat)}, c_2^{(nat)}, \ldots, c_L^{(nat)})$ and $C^{(gen)} = (c_1^{(gen)}, c_2^{(gen)}, \ldots, c_L^{(gen)})$ denote the natural and generated codon sequences, respectively. Codon-level similarity was then defined as the proportion of positions with identical codons:

$$Similarity = \frac{1}{L}\sum_{i=1} 1(c_i^{(nat)} = c_i^{(gen)}) \times 100, \quad (12)$$

where $1(\bullet)$ is the indicator function, which equals 1 if the codons at position i are identical and 0 otherwise. The resulting similarity score ranges from 0 to 100, representing the percentage of codon positions that are exactly conserved.

3.2. Jaccard Index

To quantify the overlap in codon usage between model-generated sequences and natural coding sequences, we employed a set-based similarity metric using the Jaccard index. For each coding sequence, the nucleotide sequence was first segmented into

non-overlapping codons of length three. Let $C^{(nat)} = (c_1^{(nat)}, c_2^{(nat)}, \dots)$ and $C^{(gen)} = (c_1^{(gen)}, c_2^{(gen)}, \dots)$ denote the sets of unique codons observed in the natural and generated sequences, respectively.

The Jaccard similarity between the two codon sets was defined as:

$$J = \frac{|C^{(nat)} \cap C^{(gen)}|}{|C^{(nat)} \cup C^{(gen)}|}, \tag{13}$$

where $|\bullet|$ denotes set cardinality. This metric ranges from 0 to 1, with higher values indicating greater overlap in the repertoire of codons used. Unlike position-wise identity metrics, the Jaccard index evaluates codon diversity overlap independent of positional alignment, capturing whether a model recapitulates the set of codons utilized in natural sequences rather than their exact ordering or frequency.

For each protein, Jaccard similarity was computed between the natural coding sequence and sequences generated by different codon optimization methods, including CodonTransformer, ICOR, and NatCodon.

3.3. CAI calculation

The Codon Adaptation Index (CAI) measures the overall preference of a codon sequence for optimal synonymous codons, defined as the geometric mean of normalized codon adaptiveness values:

$$CAI = (\prod_{i=1}^{L} w_i)^{1/L},\ w_i = \frac{f_{c_i}}{max_{j \in C_{\alpha_i}}(f_{c_j})}, \quad (14)$$

where L is the number of codons, f_{c_i} denotes the frequency (or adaptiveness) of codon c_i, and C_{a_i} is the set of synonymous codons encoding the same amino acid.

3.4. GC content

To assess nucleotide composition bias in coding sequences, we computed the GC content for each DNA sequence as the proportion of guanine (G) and cytosine (C) nucleotides. For a given codon sequence of length L, the GC content was defined as:

$$GC = \frac{N_G + N_C}{L} \times 100, \quad (15)$$

where N_G and N_C denote the number of guanine and cytosine nucleotides in the sequence, respectively. The resulting value is expressed as a percentage ranging from 0 to 100. All input sequences were converted to uppercase prior to computation to ensure consistent counting of nucleotide characters.

3.5. tAI calculation

To account for translational efficiency, we incorporated the tAI, which reflects codon–tRNA compatibility. For each codon c_i, a weight w_i^{tAI} is assigned based on precomputed tRNA adaptiveness values.

The sequence-level tAI is computed as the geometric mean:

$$\mathrm{tAI} = \exp\left(\frac{1}{\mathrm{L}}\sum_{\mathrm{i}=1}^{\mathrm{L}} \log \mathrm{w}_{\mathrm{i}}^{\mathrm{tAI}}\right). \quad (16)$$

3.6. %MinMax

To characterize local variations in codon optimality along codon sequences, we employed a sliding-window %MinMax metric, which quantifies the relative deviation of observed codon usage from expected synonymous codon distributions. For each codon c, four reference values were defined within its synonymous codon family. The definitions and descriptions of all reference values are provided in Supplementary Table 1.

Definition	Description
$\mathrm{f}_{\mathrm{act}}(\mathrm{c})$	The observed codon frequency (adaptiveness).
$f_{min}(c)$	The minimum frequency among synonymous codons..
$f_{min}(c)$	The maximum frequency among synonymous codons.
$f_{avg}(c)$	The expected frequency under uniform usage.

Supplementary Methods Table 2 Description of reference values for %MinMax calculation.

$f_{avg}(c)$ is biven by:

$$f_{avg}(c) = \frac{1}{|C_A|}, \quad (17)$$

where $\mathrm{C_A}$ denotes the set of synonymous codons encoding amino acid A. For each coding sequence, the nucleotide sequence was segmented into codons and analyzed

using a sliding window of fixed size $w = 18$ codons. For each window W, the average values were computed, noted as $\overline{f}_{act}$, $\overline{f}_{max}$, $\overline{f}_{min}$, $\overline{f}_{avg}$, respectively. The %MinMax value for each window was computed as a normalized deviation from the average codon usage baseline. If $\overline{f}_{act} > \overline{f}_{avg}$ (codon usage biased toward optimal codons):

$$\%Max = \frac{\overline{f}_{act} - \overline{f}_{avg}}{\overline{f}_{max} - \overline{f}_{avg}} \times 100. \quad (18)$$

If $\overline{f}_{act} \leq \overline{f}_{avg}$(codon usage biased toward non-optimal codons):

$$\%Min = \frac{\overline{f}_{avg} - \overline{f}_{act}}{\overline{f}_{avg} - \overline{f}_{min}} \times 100. \quad (19)$$

The %MinMax profile provides a position-resolved measure of codon optimality along a codon sequence. Positive regions indicate translation acceleration zones, enriched in optimal codon. Negative regions indicate translation deceleration zones, enriched in rare codons.

3.7. Kinetic distances

To quantify the similarity between model-generated and natural translation dynamics, we compared their corresponding %MinMax profiles using a position-wise distance metric. The discrepancy between model-generated and natural profiles was quantified using the mean absolute deviation:

$$D = \frac{1}{L}\sum_{i=1}^{L} |x_i^{(gen)} - x_i^{(nat)}|, \quad (20)$$

where L denotes the sequence length, $x_i^{(gen)}$ and $x_i^{(nat)}$ represent the %MinMax values at position i for the generated and natural sequences, respectively. This metric measures the average absolute difference between the two profiles at each aligned position, with lower values indicating higher similarity in local translation dynamics.

3.8. Principal component analysis

To compare codon usage–derived positional profiles across species and models, we performed principal component analysis (PCA) on %MinMax sequences. For each sequence, the original position-dependent profile was first normalized to a fixed length of 100 using linear interpolation. Specifically, profiles of varying lengths were mapped onto a uniform grid in the interval [0,1], and interpolated to 100 evenly spaced points to ensure comparability across sequences of different lengths. All resulting profiles were then assembled into a feature matrix, where each row corresponds to a sequence and each column represents a normalized positional coordinate. To eliminate scale differences across positions, the feature matrix was standardized using z-score normalization (mean-centered and variance-scaled) prior to dimensionality reduction.

PCA was subsequently applied to the standardized matrix, and the first two principal

components (PC1 and PC2) were retained for downstream visualization and analysis. These components capture the dominant modes of variation in codon usage profiles across different species and design strategies. For visualization, each point in the PCA space represents a single sequence profile, colored according to its origin (e.g., native *E. coli*, native human, or model-generated sequences). This allows for direct comparison of global distribution patterns and clustering behavior between natural and designed sequences.

3.9. Structure-conditioned codon divergence

To quantify how well designed sequences recapitulate native codon usage under structural constraints, we defined a structure-conditioned codon divergence metric based on Jensen – Shannon (JS) divergence. For a sequence set, codon usage was computed conditional on secondary structure states. Specifically, for each discrete structure token $s \in S$ and amino acid $a \in A$, we calculated the normalized codon distribution:

$$P(c|a,s) = \frac{N(c,s)}{\sum_{c' \in C(a)} N(c',s)}, \tag{21}$$

where $N(c,s)$ is the count of codon c observed in structural state s, $C(a)$ denotes the set of synonymous codons encoding amino acid a. Codons that were inconsistent with the translated amino acid sequence were discarded to ensure correctness of

codon–amino acid mapping.

For each amino acid a and structure state s, we computed the Jensen–Shannon divergence between the model-generated distribution P and the native reference distribution Q:

$$JS(P_{a,s}||Q_{a,s}) = \frac{1}{2}D_{KL}(P_{a,s}||M) + \frac{1}{2}D_{KL}(Q_{a,s}||M), \quad (22)$$

where $P_{a,s}$ is codon distribution in the candidate sequence, $Q_{a,s}$ is reference codon distribution from native data, $M = \frac{1}{2}(P_{a,s} + Q_{a,s})$. D_{KL} denotes the Kullback–Leibler divergence.

To account for the fact that codon bias varies unevenly across structural contexts, we introduced a structure-dependent weighting scheme. For each amino acid a, weights over structure states were defined as the variance of native codon usage:

$$w_{a,s} = \frac{Var(Q_{a,s})}{\sum_{s'} Var(Q_{a,s'})}. \quad (23)$$

This emphasizes structural contexts where native codon usage exhibits stronger bias.

The amino acid–level divergence is then:

$$JS_a = \sum_{s \in S} w_{a,s} \bullet JS_{a,s}, \quad (24)$$

Finally, we aggregate across amino acids using weights proportional to codon degeneracy:

$$SCCD = \sum_{a \in A} \alpha_a \bullet JS_a, \quad (25)$$

where $\alpha_a = (|C(a)| - 1)$ reflects the number of synonymous alternatives available

for amino acid a, assigning higher weight to amino acids with richer codon choice.

3.10. NatCodon Idex

The NatCodon Index (NCI) is a predictive metric for protein expression derived from structure-aware extensions of the Codon Adaptation Index (CAI), the tRNA Adaptation Index (tAI), and a single-sequence formulation of structure-conditioned codon divergence (SCCD). Below, we provide a detailed description of the formulations of structure-aware CAI, single-chain SCCD, and the final expression of NCI.

3.10.1.Structure-aware CAI

To quantify how well a coding sequence aligns with structure-dependent codon usage patterns, we first defined a position-specific weight based on the empirical frequency of codons under structural constraints. Let a codon sequence of length L (in codons) be denoted as $C = \{c_1, c_2, \ldots, c_L\}$ and the corresponding structural annotation sequence as $S = \{s_1, s_2, \ldots, s_L\}$. We constructed a reference distribution $P_{nat}(c|s)$, representing the normalized frequency of codon c under structural state s, estimated from a curated dataset of native proteins. For each position i, a structure-aware weight is assigned as $w_i^{struct} = P_{nat}(c_i|s_i)$.The overall structure-aware codon score is then defined as the geometric mean:

$$\mathrm{CAI}_{\mathrm{struct}} = \left(\prod_{i=1}^{L} w_i^{\mathrm{struct}}\right)^{1/L}. \tag{26}$$

3.10.2.Single-chain SCCD

To further capture higher-order agreement between the candidate sequence and native codon usage patterns, we computed a Jensen–Shannon (JS) similarity score over codon distributions conditioned on both amino acid identity and structural state. For each amino acid a and structural state s, we compute the Jensen–Shannon divergence described in Equation (22). We convert divergence into similarity:

$$Sim_{a,s} = 1 - JS(P_{a,s} \| Q_{a,s}). \tag{27}$$

To reflect the importance of each amino acid–structure pair, we introduce two weighting factors $w_{a,s}^{freq}$ and $w_{a,s}^{var}$, refer to frequency of occurrence in the sequence and variance of the native codon distribution, respectively. The overall JS-based score is computed as:

$$Score_{JS} = \sum_{a,s} w_{a,s}^{freq} w_{a,s}^{var} Sim_{a,s}, \tag{28}$$

3.10.3.Final NCI formulation

The final NatCodon Index integrates structure-aware optimality, translational efficiency, and distributional similarity into a unified score:

$$NCI = (NCI_{struct} \cdot tAI \cdot Score_{JS})^{1/3}, \tag{29}$$

This formulation ensures that the three components contribute equally on a logarithmic scale, balancing local structure-dependent codon preferences, global

translational efficiency, and higher-order distributional agreement.

The resulting NCI score is a single value representing the overall structural-translational harmony of a gene. The score theoretically ranges from 0 to a maximum determined by the number of synonymous codons. A lower NCI score indicates that the target sequence's codon usage more closely mirrors the native structure-specific patterns of the host, which we found to be strongly correlated with higher functional protein expression [12].

4. Statistical analysis

4.1. Cliff's δ

To quantify the effect size between two distributions without assuming normality, we used Cliff's δ. Given two samples $X = \{x_i\}_{i=1}^{m}$ and $Y = \{y_j\}_{j=1}^{n}$, Cliff's δ is defined as:

$$\delta = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{sgn}(x_i - y_j), \tag{30}$$

where:

$$\mathrm{sgn}(x_i - y_j) = \begin{cases} +1, & x_i > y_j \\ 0, & x_i = y_j \\ -1, & x_i < y_j \end{cases}, \tag{31}$$

Cliff's δ ranges from -1 to 1, where values close to 1 indicate that samples in X tend to be larger than those in Y, values near -1 indicate the opposite, and values near 0 indicate substantial overlap between the two distributions.

4.2. Wasserstein distance.

To compare full distributions beyond summary statistics, we used the first-order Wasserstein distance. For one-dimensional distributions P and Q, it is defined as:

$$W_1(P, Q) = \int_{-\infty}^{\infty} |F_p(x) - F_p(x)| dx, \tag{31}$$

where $F_p(x)$ and $F_p(x)$ are the cumulative distribution functions of P and Q, respectively. Intuitively, the Wasserstein distance measures the minimal cost of transforming one distribution into another, where cost is defined as probability mass

521 times transport distance. In the discrete case, this corresponds to the optimal transport
522 cost between empirical distributions.

523 4.3. Spearman rank correlation

524 To assess monotonic relationships between variables without assuming linearity, we
525 used Spearman's rank correlation coefficient. Given paired observations
526 $\{(x_i, y_j)\}_{i=1}^{n}$, the values are first converted to ranks $r(x_i)$ and $r(y_i)$, and the
527 coefficient is computed as:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{32}$$

528 where $d_i = r(x_i) - r(y_j)$ Spearman's ρ ranges from -1 to 1, with values close to
529 1 indicating strong monotonic agreement, and values close to -1 indicating inverse
530 monotonic relationships.

531

5. In vitro experimental validation

5.1. Selection of Experimental Subjects

We selected a diverse panel of proteins spanning a wide range of intrinsic expression levels and folding complexities, including:

1. SMURF1$_{HECT}$

We selected the SMURF1$_{HECT}$ (PDB: 1ZVD, UniProt: Q9HAU4) as a representative low-expression and difficult-to-express protein. This domain originates from a human HECT-type E3 ubiquitin ligase and thus reflects a typical eukaryotic protein expressed in a prokaryotic system. Due to the lack of appropriate folding machinery and post-translational modification pathways in *E. coli*, such proteins often exhibit reduced expression levels and poor solubility[13]. In addition, the HECT domain contains a catalytically active cysteine and requires precise conformational organization, making it highly sensitive to translation kinetics and folding conditions[14]. As a result, it is prone to misfolding and aggregation, frequently forming inclusion bodies during heterologous expression[13]. Furthermore, the codon usage of human genes differs substantially from that of *E. coli*, and the presence of rare codons can lead to ribosomal pausing, decreased translation efficiency, and impaired co-translational folding[15]. Therefore, the SMURF1$_{HECT}$ represents a stringent test case

for codon optimization, particularly under low-expression and folding-limited regimes, where optimizing synonymous codon usage can play a critical role in improving protein yield and quality.

2. DHOase

We selected the Dihydroorotase (DHOase) as a representative metabolic enzyme with moderate expression difficulty. In this study, we specifically focus on the human form of DHOase, which is part of the multifunctional CAD protein complex involved in de novo pyrimidine biosynthesis. As a eukaryotic enzyme, its heterologous expression in Escherichia coli is inherently challenging due to differences in codon usage, the absence of eukaryotic folding assistance mechanisms, and the lack of appropriate cellular context for multi-domain protein organization[16]. Although DHOase adopts a relatively conserved and stable enzymatic fold, its proper folding and activity still depend on coordinated domain interactions within the native CAD complex, which are disrupted in isolated expression systems. This results in moderate expression levels and, in some cases, reduced solubility or partial misfolding.

Compared to highly complex regulatory proteins such as E3 ligases, DHOase exhibits a more stable structural core and lower folding complexity, making it less prone to severe aggregation. However, its human origin still introduces codon bias and

translation inefficiency in E. coli, which can significantly impact protein yield and folding quality[17,18]. Therefore, DHOase occupies an intermediate regime between easy-to-express reporter proteins and highly challenging eukaryotic targets.

As such, it serves as an ideal baseline protein with moderate expression difficulty, enabling systematic comparison of codon optimization effects across different expression regimes. Improvements observed on DHOase can reflect the general effectiveness of optimization strategies beyond extreme low-expression or high-expression cases.

3. Firefly luciferase

We selected the Firefly luciferase as a canonical reporter protein and a benchmark for evaluating codon optimization strategies. Firefly luciferase is one of the most widely used reporter genes in molecular biology due to its high sensitivity, broad dynamic range, and straightforward quantification via bioluminescence assays. Its expression level can be directly and quantitatively measured, making it particularly suitable for assessing translational efficiency and protein yield.

Importantly, luciferase expression is highly sensitive to multiple sequence-level determinants, including synonymous codon usage, mRNA secondary structure, and translation elongation kinetics. Previous studies have shown that even subtle changes

586 in synonymous codons can lead to significant variations in luciferase expression,
587 primarily by affecting ribosome progression, mRNA stability, and co-translational
588 folding efficiency[17,18]. In particular, codon optimization has been demonstrated to
589 substantially enhance luciferase expression in heterologous systems, making it a
590 classical model for studying the effects of codon bias on protein production[19].
591 Furthermore, luciferase is prone to misfolding and loss of activity under suboptimal
592 translation conditions, which further amplifies its sensitivity to sequence design.
593 Therefore, firefly luciferase serves as a highly sensitive and well-established
594 benchmark protein for codon optimization. Improvements observed in luciferase
595 expression provide strong evidence for the effectiveness of sequence optimization
596 strategies, especially in capturing subtle effects related to translation dynamics and
597 mRNA structure.

598 **4. Green fluorescent protein**

599 We selected the Green fluorescent protein (GFP) as a representative highly expressed
600 and structurally stable protein to serve as an upper-bound control for expression
601 efficiency. GFP is widely recognized for its exceptional folding robustness and
602 intrinsic stability, as it can efficiently form its native β-barrel structure and mature into
603 a fluorescent state without requiring cofactors or complex post-translational

modifications. These properties enable GFP to achieve high expression levels and solubility across a variety of heterologous systems, including Escherichia coli, making it one of the most extensively used reporter proteins for evaluating gene expression and protein production.

Due to its strong tolerance to variations in translation conditions, GFP is generally less prone to aggregation or misfolding compared to more complex or unstable proteins. However, previous studies have demonstrated that its expression level can still be modulated by synonymous codon usage and mRNA sequence features, indicating that even highly optimized proteins remain sensitive to translational regulation at a finer scale[17,20]. This makes GFP particularly useful for assessing whether codon optimization methods can provide additional improvements even in already favorable expression contexts.

Therefore, GFP serves as a high-expression benchmark and upper-bound reference, enabling evaluation of whether optimization strategies yield consistent gains across the full spectrum of expression difficulty, including cases where baseline expression is already high.

5. 4-Hydroxyphenylpyruvate dioxygenase

We selected the 4-Hydroxyphenylpyruvate dioxygenase (HPPD) as a representative

622 enzyme with moderate-to-low expression levels and strong relevance to industrial
623 biotechnology applications. HPPD is a key enzyme in tyrosine catabolism, catalyzing
624 the conversion of 4-hydroxyphenylpyruvate to homogentisate, and is widely studied
625 in the context of metabolic engineering, herbicide development, and bioproduction
626 pathways. Due to its catalytic role and dependence on proper active-site configuration,
627 its functional expression requires accurate folding and coordination of metal cofactors,
628 which can be sensitive to the heterologous expression environment.
629 In recombinant systems such as Escherichia coli, HPPD often exhibits variable
630 expression levels and inconsistent solubility, depending on the host strain, expression
631 conditions, and coding sequence design. Such variability is partly attributed to
632 differences in codon usage, translation kinetics, and folding efficiency, which can
633 significantly affect enzyme yield and activity[17,18]. Compared to highly stable reporter
634 proteins, HPPD is more sensitive to sequence-level optimization, yet it is generally
635 less structurally complex than large multi-domain eukaryotic proteins, placing it in an
636 intermediate-to-challenging expression regime.
637 Therefore, HPPD serves as a representative industrial enzyme with practical
638 expression constraints, reflecting realistic scenarios where improving protein yield
639 and stability is essential for downstream applications. Its inclusion enables evaluation

of codon optimization methods in contexts that are directly relevant to biotechnology and enzyme engineering, where even moderate improvements in expression can have significant practical impact.

6. Cry1Ab toxin

We selected the Cry1Ab toxin as a representative extremely difficult-to-express protein. Cry1Ab is derived from Bacillus thuringiensis (Bt), a well-known insecticidal bacterium, and belongs to the family of crystal (Cry) toxins that exhibit high specificity against certain insect larvae. Despite its biological importance in agricultural biotechnology, Cry proteins are notoriously challenging to express in heterologous systems such as Escherichia coli.

In E. coli, Cry1Ab frequently exhibits very low soluble expression and tends to form inclusion bodies, primarily due to its large size, complex folding requirements, and incompatibility with the bacterial protein-folding environment. In addition, the gene exhibits substantial differences in GC content and codon usage bias compared to the E. coli host, which can lead to inefficient translation, ribosomal stalling, and improper co-translational folding. These factors collectively result in severely reduced functional protein yield unless extensive optimization strategies are applied, including codon optimization, fusion tags, or refolding protocols[21].

Therefore, Cry1Ab represents an extreme low-expression and high-difficulty benchmark protein, providing a stringent test case for evaluating whether codon optimization methods remain effective under severe expression bottlenecks. Its inclusion allows assessment of model robustness in scenarios where native sequence design is highly suboptimal for heterologous expression.

5.2. Gene synthesis and cloning

Codon sequences for target proteins (GeneOptimizer[22] optimized, and NatCodon designed) were synthesized and cloned into a pET-28a(+) expression vector using the NdeI and XhoI restriction sites, resulting in constructs with an N-terminal His6-tag[23]. Plasmids were transformed into chemically competent *E. coli* BL21(DE3) cells.

5.3. Protein expression and solubility analysis

To mitigate potential expression variability arising from clonal differences, overnight starter cultures were prepared by pooling 10 individual colonies into 5 mL of LB medium (50 μg/mL kanamycin) and incubating at 37°C. This starter culture was used to inoculate 250 mL of LB medium, which was grown at 37°C to an OD_{600} of 0.6-0.8. Protein expression was induced with 0.5 mM IPTG, and cultures were incubated at 16°C for 16 h. Cells were harvested by centrifugation, and normalized cell pellets (based on OD_{600}) were resuspended in lysis buffer (10 mM PBS, pH 7.4). Cells were

lysed by sonication on ice. The lysate was clarified by centrifugation (12,000 x g, 40 min, 4°C) to separate the soluble fraction (supernatant) from the insoluble fraction (pellet). The total soluble protein concentration was quantified by SDS-PAGE densitometry.

Protein samples were mixed with SDS loading buffer and heated at 95°C for 5–10 minutes to ensure complete denaturation and reduction of disulfide bonds. Equal volumes of samples were loaded for electrophoresis. Proteins were separated by SDS-PAGE using an appropriate acrylamide gel. Electrophoresis was performed at a constant voltage until the dye front reached the bottom of the gel. After electrophoresis, gels were stained using Coomassie Brilliant Blue to visualize protein bands, followed by destaining until clear background was obtained. Stained gels were imaged using a gel documentation system under standardized exposure conditions to ensure comparability across samples. Band intensities were quantified using ImageJ. The target protein band was selected, and background signal was subtracted using an adjacent blank region. A standard curve was generated using bovine serum albumin standards with known concentrations. The relationship between band intensity and protein amount was fitted (typically linear regression). The concentration of target proteins was then calculated by interpolating their band intensities on the standard

curve.

5.4. Protein purification

For functional assays, His-tagged proteins from the soluble fraction were purified by immobilized metal affinity chromatography (IMAC) using Ni-NTA resin[25].

5.5. Mutagenesis experiments

Site-directed mutagenesis was performed to introduce sequence modifications into the target genes. Mutagenic primers were synthesized by Tsingke (Beijing, China), and PCR amplification was carried out using a PCR-based site-directed mutagenesis approach. The PCR products were treated with DpnI to remove the template plasmid and then transformed into chemically competent Escherichia coli DH5 α cells. Positive clones were verified by DNA sequencing. Confirmed mutant plasmids were subsequently used for protein expression and functional analysis.

Protein concentration was determined by A280 measurement. For luciferase variants, specific activity was determined by measuring total luminescence (Promega Luciferase Assay System) at a detection wavelength of 570 nm and normalizing it to the purified protein amount. For GFP variants, specific activity was determined by measuring fluorescence (Ex: 488 nm, Em: 509 nm) and normalizing to the protein amount.

6. Model Explanation

To interpret how NatCodon utilizes structural and sequence features to guide codon selection, we employed explainable AI (XAI) techniques, focusing on SHAP (SHapley Additive exPlanations) and Integrated Gradients (IG). These methods allow quantification of feature contributions to the model's output, providing mechanistic insights into the learned structure-codon grammar.

6.1. SHAP

SHAP is a model-agnostic interpretability framework based on cooperative game theory. It quantifies the contribution of each input feature to the model's prediction by calculating Shapley values, which represent the average marginal contribution of a feature across all possible subsets of features. Formally, for a model F with input features $x_1, x_2, \ldots, x_n$ the Shapley value ϕ_i for feature x_i is computed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} [F\{S \cup \{i\}\} - F(S)]\,, \tag{33}$$

where N is the set of all features, S is a subset of features excluding x_i, and $F(S)$ denotes the model output when only features in S are present (all other features are marginalized or set to a baseline).

6.2. IG attribution

SHAP's computational complexity grows exponentially with the number of features,

making it infeasible for high-resolution per-token analysis in long codon sequences. To overcome this, we adopted Integrated Gradients (IG) as an efficient alternative that approximates SHAP values under the assumption of feature independence. IG calculates the contribution of each structural token by integrating the model's gradients along a linear interpolation path from a baseline input to the actual input:

$$IG_i = (x_i - x_i^{baseline}) \times \int_{\alpha=0}^{1} \frac{\partial F(x^{baseline} + \alpha(x - x^{baseline}))}{\partial x_i} d\alpha\,, \quad (34)$$

where x_i is the embedding of the i-th structural token, $x_i^{baseline}$ is the PAD token representing no structural information, and F denotes the predicted codon probability for that position. This formulation captures the additional contribution of structural information over a null baseline, quantifying whether a particular structural context promotes or suppresses the selection of specific codons.

6.3. Positional profiles of tAI

To characterize position-dependent translational efficiency along coding sequences, we computed positional profiles of the tRNA adaptation index (tAI). Each codon sequence was parsed into codons $\{c_1, c_2, \ldots, c_L\}$. A codon-specific weight w_i, reflecting tRNA adaptation, was assigned based on a precomputed reference table derived from tRNA abundance estimates. The tAI profile of a sequence was defined as $tAI(c_i) = w_{c_i}$. Codons not present in the reference table were assigned a

default value of $w = 1.0$. To enable comparison across sequences of different lengths, tAI profiles were normalized to a common coordinate system. Specifically, each sequence was mapped onto the interval [0,1], and its tAI values were resampled to 100 evenly spaced bins using linear interpolation:

$$\widetilde{tAI}(x) = Interp(tAI(c_i)),\ x \in [0,1], \tag{35}$$

where $\widetilde{tAI}(c_i)$ denotes the normalized positional tAI profile.

For each method and species, normalized profiles were averaged across all sequences to obtain a representative positional trend:

$$\overline{\mathrm{tAI}}(\mathrm{x}) = \frac{1}{\mathrm{N}} \sum_{\mathrm{k}=1}^{\mathrm{N}} \widetilde{\mathrm{tAI}}_{\mathrm{k}}(\mathrm{x}), \tag{36}$$

where N is the number of sequences.

Reference

1. Berman, H. M. The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).

2. Sayers, E. W. *et al.* GenBank. *Nucleic Acids Research* **47**, D94–D99 (2019).

3. Moss, M. J., Chamness, L. M. & Clark, P. L. The Effects of Codon Usage on Protein Structure and Folding. *Annual Review of Biophysics* **53**, 87–108 (2024).

4. Saunders, R. & Deane, C. M. Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Research* **38**, 6719–6728 (2010).

5. Joosten, R. P. *et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Research* **39**, D411–D419 (2011).

6. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).

7. Vaswani, A. *et al.* Attention Is All You Need. Preprint at http://arxiv.org/abs/1706.03762 (2023).

8. Ji, Z. *et al.* Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **55**, 1–38 (2023).

9. Fan, A., Lewis, M. & Dauphin, Y. Hierarchical Neural Story Generation. in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics, Melbourne, Australia, 2018).

doi:10.18653/v1/p18-1082.

10. Holtzman, A., Buys, J., Du, L., Forbes, M. & Choi, Y. The Curious Case of Neural Text Degeneration. Preprint at https://doi.org/10.48550/arXiv.1904.09751 (2020).

11. Bengio, Y., Ducharme, R. & Vincent, P. A Neural Probabilistic Language Model. in *Advances in Neural Information Processing Systems* (eds Leen, T., Dietterich, T. & Tresp, V.) vol. 13 (MIT Press, 2000).

12. Roots, C. T., Hill, A. M., Wilke, C. O. & Barrick, J. E. Codon usage modulates the relationship between the burden and yield of protein overexpression. *J Biol Eng* **19**, (2025).

13. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in Escherichia coli: advances and challenges. *Front. Microbiol.* **5**, (2014).

14. Church, G. M., Elowitz, M. B., Smolke, C. D., Voigt, C. A. & Weiss, R. Realizing the potential of synthetic biology. *Nat Rev Mol Cell Biol* **15**, 289–294 (2014).

15. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-Sequence Determinants of Gene Expression in *Escherichia coli*. *Science* **324**, 255–258 (2009).

16. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in Escherichia coli: advances and challenges. *Front. Microbiol.* **5**, (2014).

17. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-Sequence Determinants of Gene Expression in *Escherichia coli*. *Science* **324**, 255–258 (2009).

790 18. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of
791 codon bias. *Nat Rev Genet* **12**, 32–42 (2011).
792 19. Gustafsson, C., Govindarajan, S. & Minshull, J. Codon bias and heterologous protein
793 expression. *Trends in Biotechnology* **22**, 346–353 (2004).
794 20. Teixeira, L., Ferreira, Á. & Ashburner, M. The Bacterial Symbiont Wolbachia Induces
795 Resistance to RNA Viral Infections in Drosophila melanogaster. *PLoS Biol* **6**, e1000002
796 (2008).
797 21. Schnepf, E. *et al. Bacillus thuringiensis* and Its Pesticidal Crystal Proteins. *Microbiol Mol*
798 *Biol Rev* **62**, 775–806 (1998).
799 22. Raab, D., Graf, M., Notka, F., Schödl, T. & Wagner, R. The GeneOptimizer Algorithm: using
800 a sliding window approach to cope with the vast sequence space in multiparameter DNA
801 sequence optimization. *Syst Synth Biol* **4**, 215–225 (2010).
802 23. Li, L. *et al.* Expression and purification of soluble recombinant β-lactamases using
803 Escherichia coli as expression host and pET-28a as cloning vector. *Microb Cell Fact* **21**,
804 (2022).
805 24. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of
806 protein utilizing the principle of protein-dye binding. *Analytical Biochemistry* **72**, 248–254
807 (1976).

25. Hochuli, E., Döbeli, H. & Schacher, A. New metal chelate adsorbent selective for proteins and peptides containing neighbouring histidine residues. *Journal of Chromatography A* **411**, 177–184 (1987).