Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Sheng Ye

yesheng@buaa.edu.cn

Beihang University https://orcid.org/0000-0002-1097-7283

Yujie Qin

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

Shen Ge

dProtein Biotechnologies Co. Ltd., Beijing 102206, China https://orcid.org/0000-0001-8992-1006

Kaiyue Wang

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

Mengchen Pu

Beihang University https://orcid.org/0000-0001-6282-2454

Zilin Lyu

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

Zibo Cui

Beihang University

Fenghui Shi

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

Biological Sciences - Article

Keywords: Codon Sequence Design, Co-translational Folding, Generative AI, Recombinant Expression

Posted Date: July 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9654864/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

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

Keywords: Codon Sequence Design, Co-translational Folding, Generative AI, Recombinant Expression

Abstract

Protein biogenesis is a co-translational process in which nascent polypeptide folding is inextricably coupled to ribosomal elongation kinetics. However, current codon optimization strategies rely predominantly on frequency-based heuristics, decoupling translation rates from folding requirements, leading to misfolding and inclusion body formation in heterologous hosts. Here we introduce NatCodon, a generative model that reframes codon design as a structure-informed kinetic problem, embedding structural constraints directly into sequence generation to deploy less frequent codons at folding-critical positions and impose the translational pausing required for native-state acquisition. Trained on 45,126 aligned bacterial genome-proteome-structure triplets, NatCodon deciphers the structure-codon grammar embedded within natural translation systems to program co-translational folding dynamics. On cross-species benchmark datasets, NatCodon more faithfully recapitulated native coding sequence architecture than competing models, reducing divergence in translational kinetics by 96.5–97.5% and achieving an order-of-magnitude lower structure-conditioned codon divergence. Experimental validation across a panel of proteins demonstrated that NatCodon achieved 2- to 33-fold improvements in soluble protein yield over existing algorithms, including successful expression of the previously intractable $SMURF1_{HECT}$, yielding 0.57–0.61 $mg \cdot ml^{-1}$ where conventional methods produced none. Mechanistic analyses via explainable AI and site-directed mutagenesis reveal that NatCodon strategically deploys less frequent codons at structurally sensitive

sites to coordinate ribosomal pausing with nascent-chain folding—a design principle conspicuously absent in competing models. We further introduce the NatCodon Index (NCI), a structure-aware metric for predicting functional protein yield that outperforms current methods. Collectively, these advances transcend frequency-based heuristics, establishing a principled foundation for programmable protein expression that unlocks the production of therapeutically and industrially critical targets, including *de novo* AI-designed sequences previously deemed inexpressible.

1 Main

Efficient translation of genetic information into proteins underpins nearly all cellular processes and a wide range of biotechnological applications. These include the production of recombinant proteins such as therapeutic antibodies[1–4], industrial enzymes[5], and engineered transgenes for agriculture[6,7] and cell engineering[8,9], as well as emerging applications including gene therapies[10,11], nucleic acid vaccines[12], and synthetic genome design[13,14]. For many recombinant proteins, native coding sequences (CDSs) provide a starting point for optimizing expression in a target host. However, even with such guidance, designing an optimal codon sequence for a given protein to achieve reliable expression in a specific host often remains difficult[15–17]. This challenge is becoming substantially more acute in the emerging era of AI-driven protein design, where deep learning methods generate large numbers of novel protein sequences that lack corresponding genes[18,19].

Against this backdrop, codon sequences often have to be designed directly from protein sequences. However, converting protein sequences back to nucleic acid sequences is precluded by two fundamental barriers. First, no natural molecular machinery has evolved to mediate this reverse flow[20,21]; second, the inherent degeneracy of the genetic code gives rise to synonymous codons, where multiple codons encode the same amino acid[20,22].

For decades, to resolve this ambiguity, most approaches have relied on codon optimization strategies that focus on maximizing the speed of peptide elongation by replacing codons with their

most frequent synonyms in the target host[22–24]. This principle, quantified by metrics such as the Codon Adaptation Index (CAI)[22] and the tRNA Adaptation Index (tAI)[25], is embedded in nearly all conventional codon optimization algorithms. The advent of AI has catalyzed a new wave of computational approaches for codon sequence design, moving beyond simplistic frequency-based heuristics toward the recognition of high-dimensional sequence patterns. Early efforts employed recurrent neural networks such as Long Short-Term Memory models to capture codon dependencies along CDSs[26,27]. More recently, Transformer-based architectures have been adopted due to their superior capacity for modeling long-range sequence relationships and scalable training[28–30]. In parallel, several studies have attempted to incorporate additional regulatory signals, including mRNA structural features, to better constrain the vast design space of synonymous codons[31]. Despite these advances, existing pipelines still model codon sequence design primarily as a seq-to-seq mapping problem between amino acid sequences and CDSs, discarding rich structural information of proteins.

This limitation becomes particularly important in light of the dynamic nature of translation. The degeneracy of the genetic code creates an enormous codon sequence search space relative to the amino acid sequence, leaving models under-constrained when relying solely on protein sequences. Moreover, existing models often treat translation as a static process, overlooking the precise regulatory role of translation rhythm in coordinating co-translational folding (CTF)[32–34]. CTF refers to the folding process of a newly synthesized peptide while still tethered to the ribosome, which is

highly dependent on a precisely choreographed translational rhythm, shaped by the integration of genomic (CDS), structural (protein structure), and translatomic (ribosome occupancy) signals[35,36]. Strategically placed less-frequent codons—identified via ribosome profiling as sites of transient translational pausing—are critical for the proper formation of complex structural motifs, domain boundaries, and active sites[34,37,38]. As a result, codon sequence designs generated by current methods still fail to fully capture the structure-codon grammar embedded in natural translation systems.

Here, we first confirmed that synonymous codon usage in native bacterial proteins is closely associated with local protein structural context. To leverage this relationship, we developed NatCodon, a protein-structure-informed generative model that models codon sequence design by integrating multi-omics datasets of native bacteria, including genomic sequences, proteomic context and structural representations of proteins. By learning the relationships between protein structures and host-specific codon usage patterns, NatCodon enables codon sequence design while coordinating translation rhythm with the folding requirements of protein. Across diverse datasets, NatCodon consistently outperforms existing codon sequence design models, particularly improving the preservation of native translation kinetics while reducing structure-conditioned codon divergence. Experimental validation further demonstrates NatCodon is able to improve the production of correctly folded heterologous proteins. Moreover, model interpretation analyses reveal how NatCodon strategically allocates synonymous codons at key structural transitions to coordinate

translation with CTF. In addition, we introduce the NatCodon Index (NCI), a structure-aware metric that quantifies the concordance between codon usage and structural context within the host proteome. Together, these advances provide a structure-informed framework for codon sequence design, founded on the principle that synonymous codon usage is intrinsically coupled to target protein structural context, and demonstrate that leveraging this coupling can better orchestrate translation dynamics with co-translational protein folding. The NatCodon programs have been made freely accessible to the community through https://dprotein.cn/codon/.

2 Results

2.1 Native Codon Selection is Coupled to Protein Structure

Substantial evidence suggests that CTF is intrinsically linked to translational kinetics, where the strategic selection of synonymous codons modulates ribosomal elongation speed to ensure proteome integrity[23,36,38] (Fig. 1a). Building upon this framework, we proposed a model-NatCodon-designed to engineer codon sequences by integrating both primary amino acid sequences and local structural information. To operationalize this approach, we defined 9 discrete structural states (tokens) derived from protein secondary structures, including α-helix (H), isolated β-bridge (B), extended strand (E), 3_{10}-helix (G), π-helix (I), hydrogen-bonded turn (T), bend (S), flexible loop (F), unmodeled region (X). To validate the efficacy of these structural descriptors, we performed an extensive analysis of 45,126 non-redundant bacterial proteins to quantify the dependency between these tokens and

synonymous codon usage.

Initial assessment via Mutual Information (MI) revealed a proteome-wide association of 0.068 ($p < 0.001$), establishing a non-random, global dependency between codon identity and structural context. This confirms that synonymous codon selection is not structurally neutral. At the single-residue level, χ^2 tests followed by Cramér's V coefficients ($V \approx 0.02–0.055$) demonstrated that this coupling is pervasive across the proteome (Fig. 1b). Notably, residues critical for fold nucleation, such as Cysteine[39] and Tyrosine[40], manifested the strongest structural dependency ($V \approx 0.05$). This suggests that evolutionary pressure for kinetic regulation is intensified at pivotal folding nodes to prevent misfolding.

The biological directionality of this coupling became evident when analyzing codon enrichment against genomic backgrounds (Fig. 1c). Furthermore, by correlating codon enrichment patterns with cellular tRNA abundance indices, the mechanistic directionality of this relationship was further elucidated (Fig. 1d). Our results indicate that for the majority of structural elements, codon preference is associated with tRNA availability. Specifically, π-helix and isolated β-bridge—motifs frequently associated with the structural core [41–43]—exhibited significant positive correlations with tRNA abundance (Spearman's $r \approx 0.33$, $p < 0.05$), suggesting an evolutionary preference for accelerated elongation to minimize the exposure of hydrophobic intermediates. Similar positive trends were observed in bends and α-helix (Spearman's $r \approx 0.22–0.24$, $p < 0.1$). Conversely,

unmodeled regions and flexible loops exhibited significant and marginal negative associations, respectively (Spearman's r = -0.56, $p \ll 0.001$; Spearman's r = -0.22, $p < 0.1$), indicating a preferential recruitment of low-abundance slow codons that may provide the requisite latency for the structural maturation of preceding domains (Fig. 1a).

Collectively, these findings reveal a spatially resolved translational program embedded within the CDSs, where codon choice is tuned to the structural demands of the nascent polypeptides. This empirical structure-codon code provides the necessary biological constraints to move beyond heuristic optimization toward a structure-informed design framework.

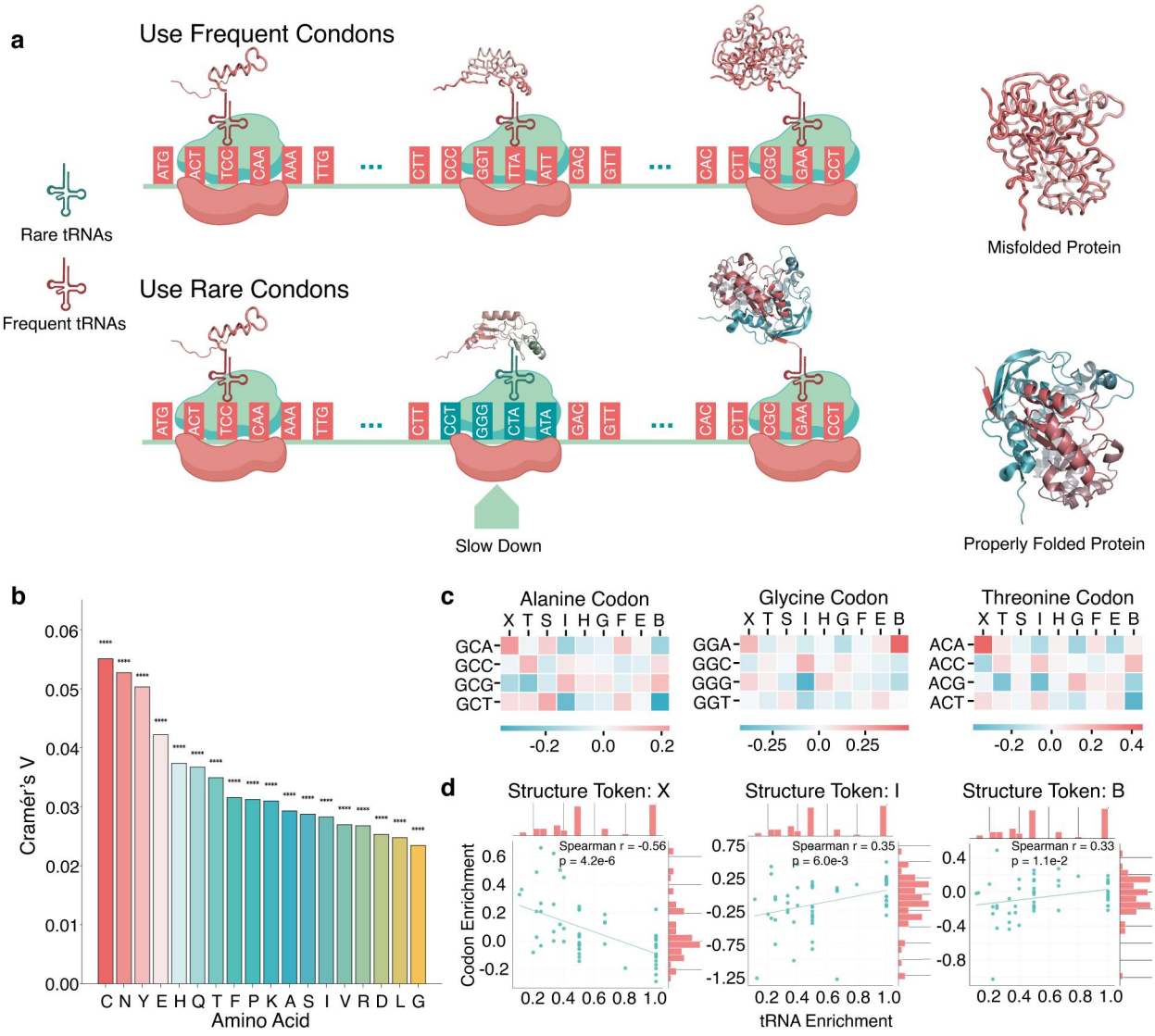


Fig. 1. **Co-evolutionary coupling between synonymous codon usage and local protein structure and its role in co-translational folding**. **a,** Schematic illustrating how codon-dependent translation kinetics modulate protein folding. Excessive use of optimal codons accelerates ribosomal elongation, limiting the time available for nascent chain folding and increasing the risk of misfolding. In contrast, strategic insertion of rare codons (dark green) induces local ribosome pausing, providing a necessary temporal window for structural maturation and promoting correct folding. **b,** Strength of association between codon usage and structural context at the amino acid level. Bar plots show Cramér's V

values quantifying the association between synonymous codon usage for each of the 20 amino acids and their distribution across 9 structural categories. **c,** Structure-dependent enrichment patterns of synonymous codons. Heatmap illustrating significant enrichment (red) or depletion (blue) of specific synonymous codons across different structural contexts. **d,** Correlation between codon enrichment and intracellular tRNA abundance. Scatter plots showing the relationship between structure-specific codon enrichment and the corresponding tRNA abundance index.

2.2 Overview of NatCodon

Although statistical evidence confirms a robust coupling between protein structure and codon usage, the multidimensionality of this relationship exceeds the capacity of existing methods. To capture these intricate, non-linear dependencies, we developed NatCodon—a protein-structure-informed codon sequence design model to orchestrate CTF for efficient heterologous protein expression. NatCodon adopts a decoder-only Transformer architecture inspired by the LLaMA[44] framework (Fig. 2a). Amino acid identity and discrete protein structure are first embedded into a unified representation, which is integrated with rotary positional embeddings (RoPE) to enable efficient modeling of long-range dependencies across the CDSs. Stacked self-attention decoder layers then predict codon probability distributions conditioned on this structure-aware context. NatCodon reframes codon sequence design as a structure-conditioned generative modeling problem rather than a frequency-based heuristic task.

Complementing this architecture, a key innovation of NatCodon lies in its biologically constrained decoding strategy (Fig. 2b). A "Codon Gate" module applies an amino-acid–conditioned hard mask to the output logits at each position, ensuring strict translational fidelity. To further enable tunable diversity of generated codon sequences, a Focus Factor (ϕ), analogous to inverse temperature scaling, modulates the sharpness of the gated probability distribution, providing controllable trade-offs between deterministic optimization and sequence diversity.

To ground the model in biological reality, NatCodon was trained on a non-redundant dataset pairing the native *E. coli* proteome with its endogenous coding sequences. Detailed descriptions of model architecture, training procedures and dataset curation are provided in the Supplementary Methods 1.

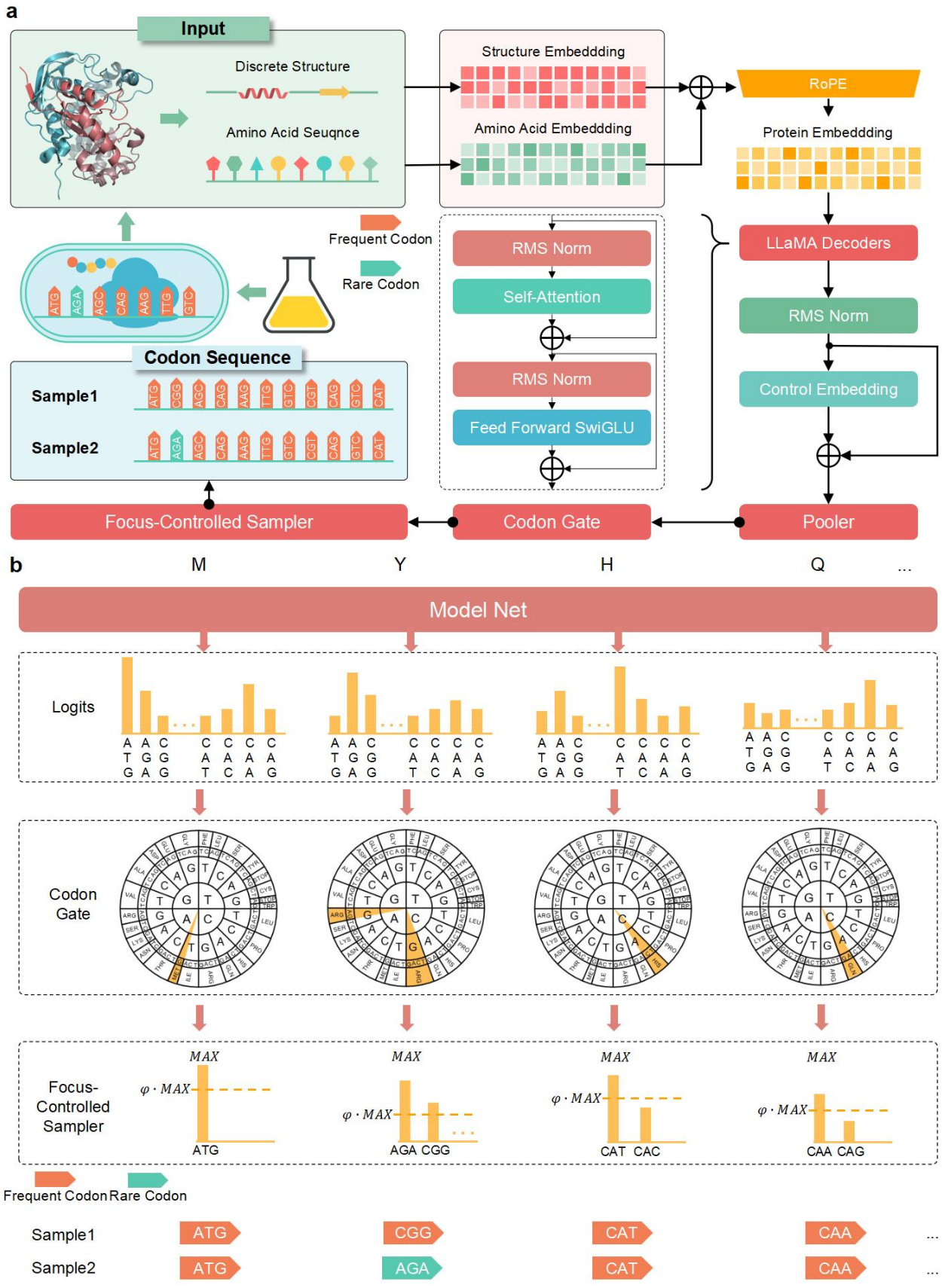

a
Input
Discrete Structure
Amino Acid Seuqnce
Structure Embeddding
Amino Acid Embeddding
RoPE
Protein Embeddding
Frequent Codon
Rare Codon
RMS Norm
Self-Attention
RMS Norm
Feed Forward SwiGLU
LLaMA Decoders
RMS Norm
Control Embedding
Codon Sequence
Sample1
Sample2
Focus-Controlled Sampler
Codon Gate
Pooler
b
M
Y
H
Q
...
Model Net
Logits
Codon Gate
Focus-Controlled Sampler
MAX
ATG
AGA CGG
CAT CAC
CAA CAG
Frequent Codon
Rare Codon
Sample1
Sample2
ATG
CGG
CAT
CAA
ATG
AGA
CAT
CAA

Fig. 2 **Overview of the NatCodon architecture and its biologically constrained decoding strategy**. **a,** End-to-end pipeline of the NatCodon model. The framework adopts a decoder-only Transformer architecture based on LLaMA. **Input and embedding**: The model jointly encodes amino acid sequences and discretized structural tokens (e.g., α-helices and β-sheets). These features are integrated via amino acid and structure embeddings, and combined with rotary positional encoding (RoPE) to capture long-range dependencies within codon sequences. **Backbone architecture**: The core consists of stacked LLaMA decoder layers, incorporating RMS normalization, self-attention, and SwiGLU activations. **Control and generation**: Following a pooling layer, model outputs are processed through a codon gate module and a focus-controlled sampler to generate diverse candidates for downstream Recombinant expression validation. **b,** Biologically constrained decoding and sampling mechanism. The model transforms raw probability outputs into biologically valid codon selections. **Logits prediction**: For each amino acid (e.g., M, Y, H, Q), the network predicts logits over all 64 codons. **Codon gate**: This module acts as a hard mask that filters out codons incompatible with the target amino acid, thereby ensuring translational fidelity. **Focus-controlled sampler**: A tunable focus parameter (ϕ), analogous to inverse temperature scaling, modulates the sharpness of the output distribution. Higher ϕ values bias selection toward the most probable codons (e.g., Sample 1), whereas adjusting ϕ increases sequence diversity (e.g., selecting AGA at a arginine position in Sample 2), enabling a balance between deterministic optimization and

codon diversity.

2.3 Cross-species benchmarking of NatCodon

To assess whether NatCodon internalizes native coding principles beyond organism-specific codon preferences, we conducted a cross-species benchmark across four phylogenetically distant organisms: *E. coli* (2,507 proteins), human (1,331 proteins), plant (845 proteins), and virus (479 proteins) CDSs. All test datasets were strictly non-overlapping with the training and validating data of NatCodon. For each species, sequences generated by NatCodon were compared with those produced by ICOR and CodonTransformer, using the corresponding native CDSs as references. Our evaluation framework spanned three hierarchical dimensions: global compositional stability, preservation of translation kinetics landscapes, and structural codon fidelity.

On the *E. coli* dataset, NatCodon showed strong recovery of native coding patterns, achieving large effect sizes across similarity and Jaccard metrics (Cliff's δ = 0.80-0.90, $p < 10^{-16}$, Fig.3a-b), indicating substantially higher concordance with native sequences.

To assess cross-species optimization performance at the compositional level, we quantified global nucleotide properties using the Codon Adaptation Index (CAI; Fig. 3c) and GC content (Fig. 3e). Since the native *E. coli* CDSs distribution reflects sequences empirically optimized for the host's translational and regulatory machinery, any deviation from this baseline serves as a proxy for the disruption of host-adapted coding constraints. Our analysis reveals that NatCodon maintains minimal

CAI drift (Wasserstein distance < 0.03; Fig. 3d), remaining close to the natural compositional landscape. In contrast, CodonTransformer and ICOR caused substantially larger deviations (Wasserstein distance ≈ 0.2), representing a 6–7-fold greater displacement owing to their excessive reliance on codons with high relative adaptiveness. At the GC content level, NatCodon preserved near-native distribution on the *E. coli* dataset (Wasserstein distance = 0.19; Fig. 3f), whereas competing models showed pronounced shifts (Wasserstein distance > 5). Across heterologous datasets, NatCodon reduced GC disparity for plant sequences (8.07 → 4.20) and viral sequences (9.70 → 6.02). In contrast, competing models frequently induced larger compositional shifts. These findings indicate that NatCodon maintains proximity to the native functional coding space while adaptively modulating nucleotide composition across species, rather than imposing uniform or excessive codon enrichment.

At the kinetic level, translational dynamics were quantified using %MinMax profiles, which approximate position-specific elongation rates along the codon sequences. NatCodon showed consistently large reductions in divergence, outperforming CodonTransformer and ICOR in 96.5% and 97.5% of cases, respectively (Cliff's δ = 0.93-0.95; $p < 10^{-16}$; Fig. 3g). Biologically, reduced %MinMax divergence indicates preservation of endogenous elongation-rate fluctuations, including native pause-site positioning and local translation speed transitions that are shaped by host-specific tRNA availability and CTF requirements. Principal component analysis further showed

234 that NatCodon-generated sequences largely remained within the native kinetic manifold across
235 human, plant, and viral datasets (Fig. 3h-j), whereas competing models displaced sequences toward
236 non-native kinetic regimes. Maintenance of this native-like kinetic architecture suggests that
237 NatCodon retains the temporal structure of translation, potentially supporting proper folding
238 trajectories and avoiding artificial smoothing of elongation dynamics caused by over-optimization.

239 We quantified structure-codon fidelity using a structure-conditioned codon divergence (SCCD;
240 Fig. 3k; see Supplementary Methods 3.9) metric based on weighted Jensen–Shannon (JS) divergence.
241 This metric captures the extent to which synonymous codon preferences remain coupled to structure
242 context, reflecting preservation of CTF signals. Lower values indicate preservation of native
243 coupling between codon choice and local structural environment. Across all species datasets,
244 NatCodon consistently exhibited the smallest SCCD relative to competing models (<0.8; Fig. 3l) and
245 in non-*E. coli* datasets, shifted codon-structure distributions toward the host reference. Biologically,
246 reduced JS divergence indicates preservation of structure-dependent synonymous codon biases that
247 are thought to coordinate local translation speed with emerging secondary-structure formation.
248 Because native structure-codon coupling reflects evolutionary tuning of co-translational folding
249 dynamics, maintenance of this relationship suggests that NatCodon retains regulatory information
250 embedded within coding sequences rather than disrupting structural translation cues through uniform
251 codon replacement. These results imply that NatCodon preserves not only global compositional

252 balance but also the fine-grained linkage between local structural context and codon choice, thereby
253 maintaining biologically meaningful folding constraints encoded at the nucleotide level.

254 In summary, cross-species benchmarking reveals that NatCodon achieves multi-scale host
255 compatibility: maintaining compositional equilibrium, preserving translational kinetics landscapes,
256 and safeguarding structure-codon coupling. Rather than maximizing codon frequency in isolation,
257 NatCodon retains higher-order organizational constraints embedded in natural CDSs. This
258 multi-scale fidelity distinguishes NatCodon from existing AI-based codon optimization frameworks
259 and underscores its potential as a structure-aware, evolution-consistent codon sequence design
260 strategy.

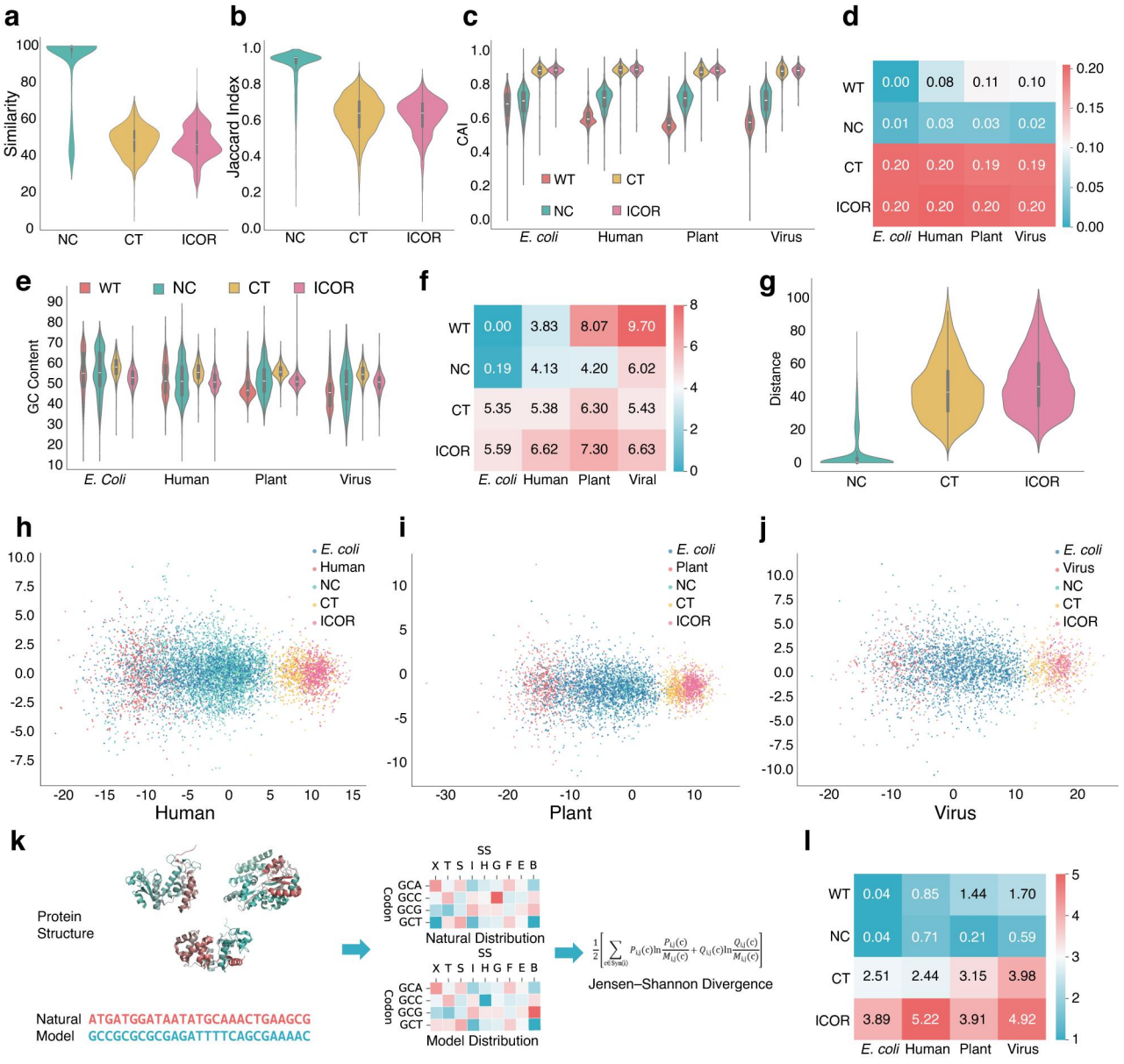


Fig. 3 **Multidimensional benchmarking of NatCodon (NC) across species**. Performance of NatCodon is compared with CodonTransformer (CT), ICOR, and wild-type (WT) sequences across four taxa (*E. coli*, human, plant, and viral datasets). **a,b,** Sequence recovery in *E. coli*. Similarity **(a)** and Jaccard index **(b)** between model-generated sequences and native CDSs. **c–f,** Global sequence features (CAI and GC content). **c,e,** Distributions of codon adaptation index (CAI) and GC content across species. **d,f,** Wasserstein distances relative to WT for CAI and GC content. **g–j,** Translational

kinetics landscapes. **g,** Kinetic distances derived from %MinMax profiles. **h–j,** Principal component analysis (PCA) of kinetic profiles for human **(h)**, plant **(i)**, and viral **(j)** datasets. **k,** Schematic of SCCD calculation based on weighted JS divergence between native and generated distributions. **l,** SCCD heatmap across species.

2.4 NatCodon enables soluble expression of intractable proteins

While *in silico* metrics provide valuable initial proxies, the ultimate validation of a codon sequence design strategy lies in its capacity to enhance protein production. We therefore evaluated NatCodon on a panel of heterologous proteins with diverse origins. To obtain robust estimates of expression performance, ten well-grown colonies were selected from each transformation plate and screened for soluble protein production. To minimize variability arising from differences in cell density, cultures were normalized based on OD600 prior to harvesting, ensuring equivalent biomass was used for downstream soluble protein analysis (Fig. 4a).

Across five tested proteins, NatCodon-designed sequences (NC) consistently improved soluble expression compared with sequences optimized by GeneOptimizer (OP). Notably, several proteins that exhibited negligible expression under OP optimization became readily detectable when redesigned by NatCodon. For example, HECT domain of SMURF1 ($SMURF1_{HECT}$, PDB: 1ZVD, UniProt: Q9HAU4) showed no detectable soluble expression with OP optimization, as confirmed by Western blot analysis of the soluble fraction, whereas NatCodon enabled strong production (0.57

286 $mg \cdot ml^{-1}$; Fig. 4b; Supplementary Information Figures 2). Similar improvements were observed for
287 DHOase (PDB: 6HFN, UniProt: P00374), Luciferase (PDB: 5DV9, UniProt: P08659), GFP (PDB:
288 1GFL, UniProt: P42212) and HPPD (PDB: 5EC3, UniProt: P32754), with NatCodon increasing
289 expression levels by approximately 2.25-33 fold(Fig. 4b,c). Across all tested proteins, NC
290 consistently outperformed OP, yielding a large overall effect size (Cliff's $\delta = 0.68$, $p = 0.031$;
291 *two-sided Wilcoxon signed-rank test*). The rationale for selecting these proteins is provided in
292 Supplementary Information Section 5.1.

293 We observed that for several proteins, experimentally measured expression levels did not
294 correlate well with CAI-based predictions (Fig. 4d), indicating that CAI alone may not fully capture
295 translational efficiency in these systems. To assess the robustness of these observations, we
296 performed three independent biological replicates for four representative proteins (DHOase,
297 $SMURF1_{HECT}$, GFP, and luciferase). Across all replicates, NC consistently outperformed OP (Fig. 4e).
298 $SMURF1_{HECT}$ remained undetectable in every OP experiment, whereas NatCodon enabled stable and
299 substantial soluble expression in all trials (0.57–0.61 $mg \cdot ml^{-1}$). Similarly, DHOase exhibited only
300 minimal expression under OP optimization (0.00–0.01 $mg \cdot ml^{-1}$), but showed significantly increased
301 soluble production under NatCodon design (0.13–0.33 $mg \cdot ml^{-1}$), with comparable trends observed
302 for GFP and Luciferase (Fig. 4f). Collectively, these findings suggest that CAI alone is insufficient to
303 fully explain translational outcomes in these contexts. Importantly, NatCodon enables efficient

soluble protein expression through codon reprogramming alone, without altering the amino acid sequence, highlighting the functional impact of optimized synonymous codon selection.

To determine whether increased soluble expression corresponded to functional protein production, we quantified Luciferase luminescence and GFP fluorescence. Luciferase functional expression, represented by luminescence, increased by 1.5-3.3-fold under NatCodon design (Fig. 4f). GFP fluorescence measurements showed a similar trend, with NatCodon constructs producing higher functional output in independent experiments. These results indicate that NatCodon improves not only protein yield but also the production of functional proteins.

To investigate the role of structural information, we performed an ablation experiment by removing structural features during sequence design (NC-NS). While the structure-free model still outperformed OP in several cases, its performance was consistently inferior to the full NatCodon model. For instance, $SMURF1_{HECT}$ expression dropped from 0.57–0.61 $mg \cdot ml^{-1}$ under the full model to 0.10–0.15 $mg \cdot ml^{-1}$ without structural inputs, and Luciferase decreased from 0.13-0.24 $mg \cdot ml^{-1}$ to 0.00-0.03 $mg \cdot ml^{-1}$. Similar trends were observed for DHOase (Fig. 4g). These results indicate that structural context contributes substantially to the improved expression enabled by NatCodon.

Finally, we tested the system on an additional protein, Cry1Ab (UniProt: A0A7D8ZIV8), which further confirmed the importance of structural guidance: while the structure-aware NatCodon design achieved detectable expression (0.012-0.023 $mg \cdot ml^{-1}$), the structure-free variant failed to produce

measurable protein.

Together, these results demonstrate that NatCodon can substantially enhance soluble protein production through synonymous codon reprogramming alone. By incorporating structural context during sequence design, the model enables robust expression improvements and may rescue proteins that remain intractable under conventional optimization strategies.

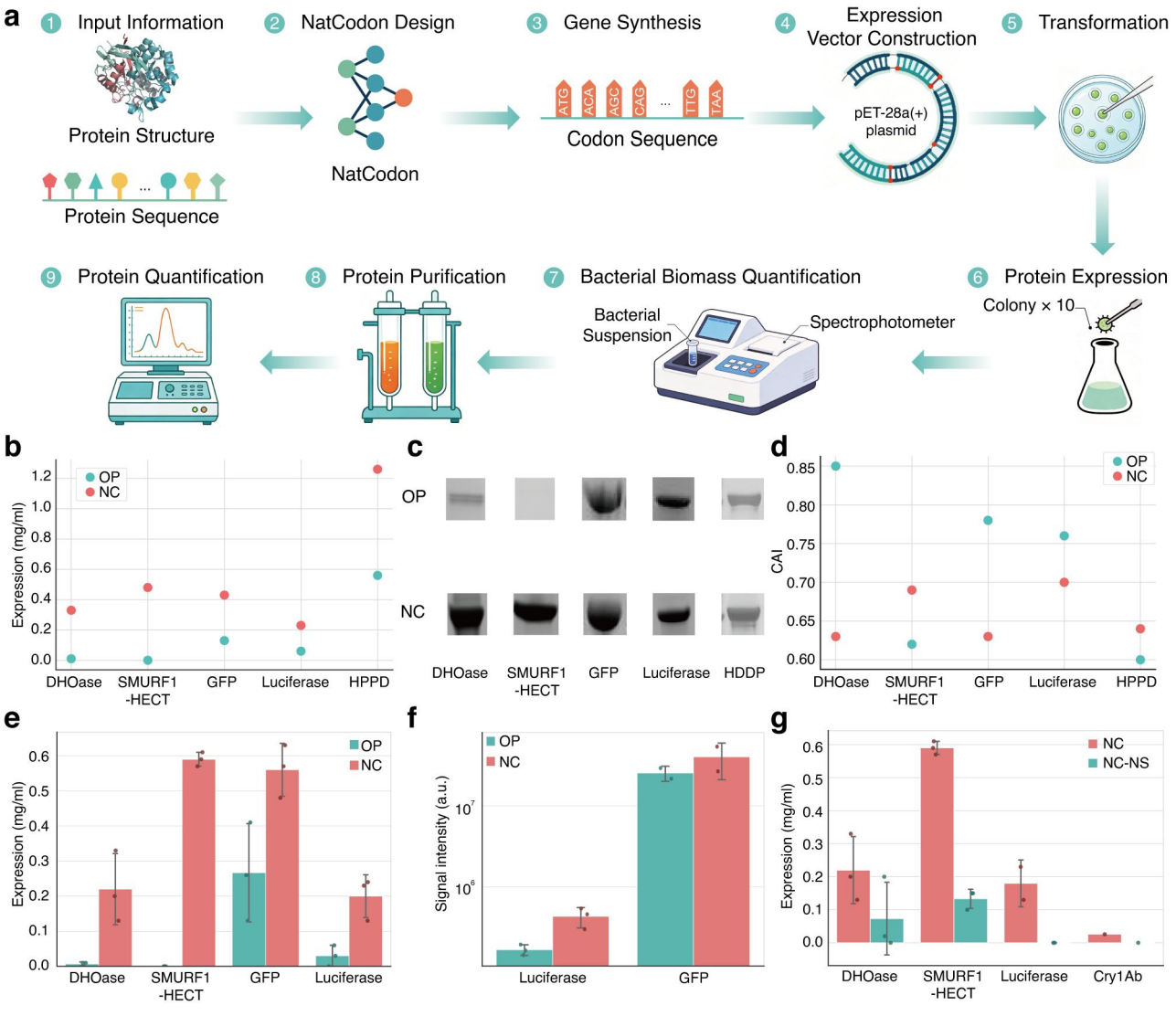


Fig. 4 **NatCodon enhances solubility and functional expression of difficult-to-express proteins in experimental validation**. **a,** Schematic of the experimental workflow, including computational

design using NatCodon (NC) and GeneOptimizer (OP), gene synthesis and cloning into the pET-28a(+) vector, transformation into *E. coli* BL21(DE3), induction, cell lysis and centrifugation, and protein quantification by densitometry, SDS-PAGE, and functional assays. **b,c**, Screening of five heterologous proteins. **b,** Scatter plot showing soluble expression levels of NC (red) and OP (teal) designs for DHOase, $SMURF1_{HECT}$, GFP, luciferase, and HPPD. **c,** Representative SDS-PAGE gels. **d,** Relationship between expression level and codon adaptation index (CAI). **e,f,** Reproducibility and functional assays. **e,** Soluble expression levels of four representative proteins across three independent biological replicates. **f,** Functional activity measurements for Luciferase (luminescence) and GFP (fluorescence). **g**, Ablation of structural information. Comparison between the full NatCodon model (NC) and a structure-agnostic variant (NC-NS) across multiple proteins.

2.5 Structure-Codon Grammar and Translation Kinetics Encoded by NatCodon

The superior functional output of NatCodon-designed proteins suggests the model does more than simply optimize codon frequency; it appears to choreograph the entire process of protein biogenesis. We hypothesized that NatCodon achieves this by encoding the structure-codon grammar of CTF: (i) introducing localized translational pauses at structurally sensitive regions to spare time for nascent-chain folding; (ii) maintaining efficient translation within stable structural elements; (iii) implementing a global N-terminal translational ramp to regulate ribosome traffic.

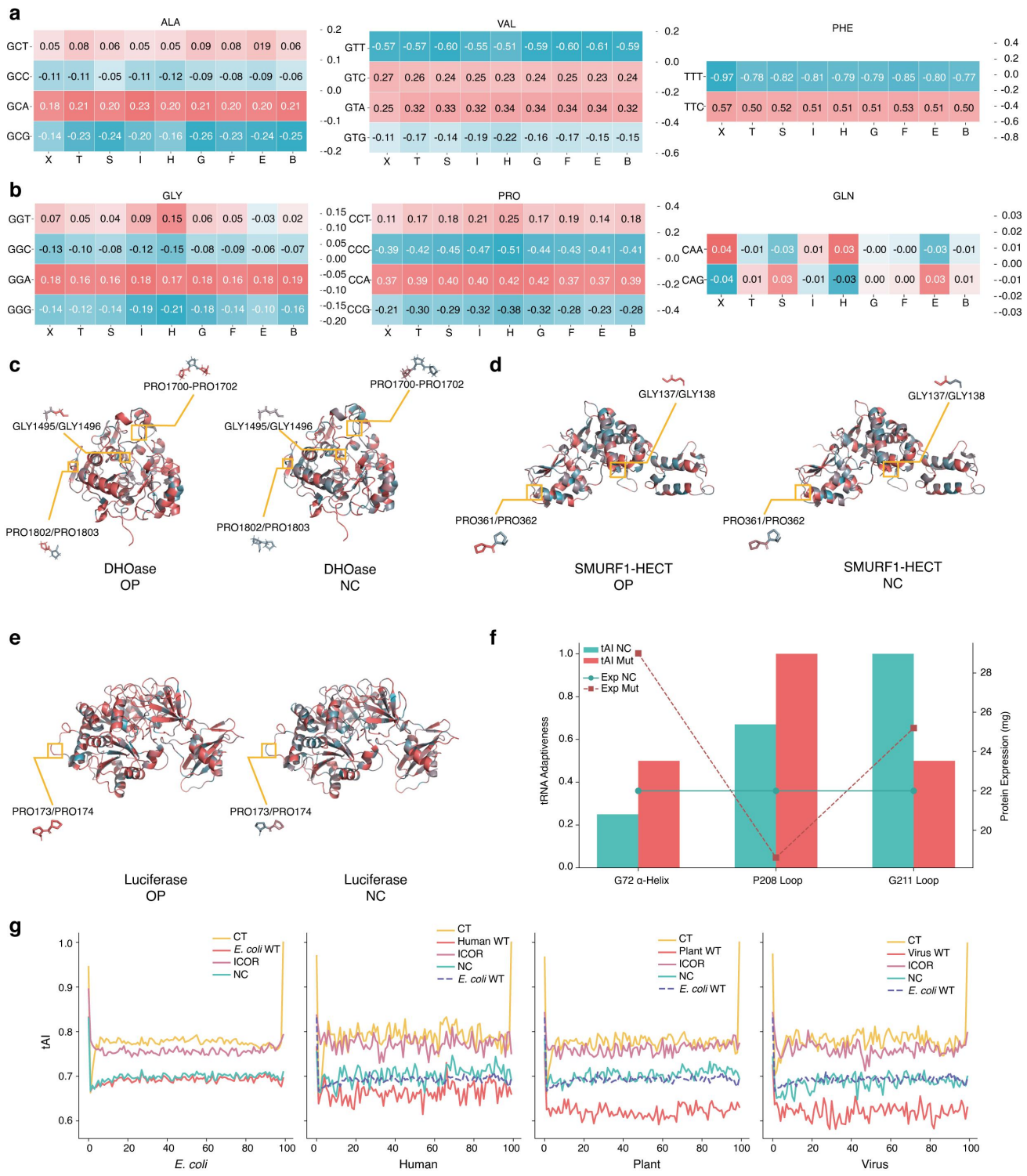
To test these hypotheses, we first employed explainable AI (XAI) to decipher the learned relationship between structural grammar and translation kinetics. SHAP analysis revealed that NatCodon assigns nearly equal weight to structural features (0.058) and amino acid identity (0.057), suggesting an inherent optimization balance between translational efficiency and folding constraints. Integrated Gradients (IG) analysis further unmasked a sophisticated codon grammar: while the model selects high-tAI codons to accelerate synthesis of rigid structural cores (e.g., Ala, Val, Phe), it systematically favors lower-tAI codons at flexible loops and domain boundaries (e.g., Gly, Pro, Ser) to introduce strategic pauses (Fig. 5a–b, Extended Data Fig. 2). This sensitivity is exemplified by Glutamine (Gln) selection, where high-tAI CAG is prioritized for α-helices, whereas lower-tAI CAA is reserved for complex loop architectures (Fig. 5c), mirroring the structure-dependent enrichment patterns observed in native sequences (Section 2.1). This abstract grammar becomes physically apparent when mapping translation kinetics onto 3D structures. In contrast to OP, which accelerates translation indiscriminately, NC strategically places slow codons at pivotal loop junctions across diverse proteins, including DHOase (Gly18/19, Pro240-242), $SMURF1_{HECT}$ (Gly136/137, Pro360/361), and Luciferase (Gly338/340; Fig. 5d–g), potentially mitigating misfolding risks at these critical junctures.

The functional necessity of these programmed pauses was confirmed through reverse mutagenesis of NatCodon-designed DHOase. In stable α-helix regions, increasing the tAI (G72:

GGA to GGT) further boosted expression from 22 mg to 29 mg, confirming the benefit of high-speed codons in rigid units. Conversely, at loop-embedded sites (P208, G211), increasing translation speed (P208: CCA to CCG) reduced soluble yield to 18.6 mg, while further decreasing speed (G211: GGC to GGT) actually enhanced expression to 25.2 mg (Fig. 5d). These results provide direct evidence that NatCodon's use of sub-optimal codons in flexible regions is a deliberate strategy to maximize soluble yield by providing a timing window for CTF.

Beyond local kinetics, NatCodon captures macro-scale dynamics by implementing a 5'-terminal "translational ramp"—a feature characterized by a 10–20% initial segment of slower codons to prevent ribosome collisions and facilitate early folding intermediates. While native and NatCodon-designed sequences exhibit this translational ramp, other models such as CodonTransformer and ICOR generate fast-start sequences that disrupt this natural gradient (Fig. 5g). Notably, NatCodon effectively recuperates this 5'-to-3' acceleration gradient when re-engineering heterologous genes (e.g., virus, plant, and human) for *E. coli*, demonstrating its ability to adapt foreign CDSs to host-specific biophysical requirements.

Together, these findings indicate that NatCodon achieves superior protein biogenesis by synchronizing local elongation rates with the structural birth of the nascent polypeptide, effectively shifting the optimization focus from global frequency to site-specific kinetic choreography.



383

384 Fig. 5 **Structure-codon grammar and translational kinetics encoded by NatCodon (NC). a,b,**

385 Structure-codon preferences inferred by explainable AI. Heatmaps derived from Integrated Gradients

386 show codon usage preferences for representative amino acids (e.g., Ala, Val, Phe, Gly, Pro, Gln)

387 across structural tokens (X, T, S, I, H, G, F, E, B). **c–e,** Mapping of codon-dependent kinetics onto
388 protein structures. Key residues are highlighted on the 3D structures of DHOase **(c)**, $SMURF1_{HECT}$
389 **(d),** and Luciferase **(e),** comparing NC and GeneOptimizer (OP) designs. **f,** Site-directed mutagenesis
390 of DHOase. Expression levels are measured following codon substitutions at selected positions in
391 α-helical (G72) and loop regions (P208, G211). **g,** Positional profiles of tAI along coding sequences.
392 Average tAI values across the first 100 codons are shown for sequences generated by different
393 models and compared with native *E. coli*.

394 ## 2.6 The NCI metric reliably predicts functional protein yield

395 This newfound understanding of the structure-codon grammar necessitates a fundamental
396 re-evaluation of how codon sequence optimality is quantified. For decades, the field has leaned
397 heavily on the 'faster is better' heuristic, exemplified by metrics such as CAI and tAI. However, if
398 protein fitness is indeed a property of precise kinetic pacing rather than maximal elongation speed,
399 then metrics based solely on codon usage frequency become fundamentally decoupled from the
400 biophysical constraints of CTF. Our data demonstrate that this theoretical misalignment has profound
401 practical consequences, as optimizing for high CAI paradoxically compromise protein soluble
402 expression (Section 2.4).

403 To address this limitation, we developed the NatCodon Index (NCI), a structure-aware metric
404 that evaluates how well a codon sequence aligns with its protein folding trajectory. NCI captures the

405 host's native structure-codon grammar through a dual-scale approach. At the local level, it introduces
406 structure-specific adaptiveness coefficients to account for the structural context of each codon. At the
407 global level, it employs JS divergence to ensure the entire sequence conforms to the systemic
408 patterns of the host proteome. Consequently, a higher NCI score indicates a closer adherence to these
409 evolutionarily optimized patterns, ensuring that synonymous designs maintain the 'translational logic'
410 essential for native-like protein folding.

411 To validate the robustness of NCI, we analyzed two synonymous mutation libraries measuring
412 fluorescence intensity in GFP and mRFP. In the mRFP dataset, NCI achieved the highest correlation
413 with experimental output among all evaluated metrics (Spearman's r: NCI = 0.485 vs. CAI = 0.470,
414 tAI = 0.457; Fig. 6a). Similarly, in the GFP library, NCI consistently outperformed traditional metrics
415 (Spearman's r NCI= 0.471 vs. CAI = 0.440, tAI = 0.422, Encodon = 0.230; Fig. 6b). While the
416 RNA-based large language model Encodon exhibited a higher correlation in the mRFP dataset
417 (Spearman's r = 0.731), its performance is contingent upon supervised fitting to target-specific
418 experimental data. In contrast, NCI operates independently of prior experimental benchmarks,
419 offering a more versatile and generalizable framework.

420 To test the predictive power of NCI for complex targets, we pitted it against CAI, tAI and
421 Encodon across a panel of six challenging proteins (including DHOase, $SMURF1_{HECT}$). While the
422 performance of CAI, tAI and Encodon showed higher variability across our protein panel—due to

the inherent complexity of non-homologous soluble expression—NCI exhibited a strong and consistent positive correlation across all targets (Fig. 6c). A higher NCI score, signifying stronger adherence to native structure-specific codon patterns, is positively correlated with higher soluble yield.

Crucially, high soluble expression does not guarantee functional integrity, as accelerated translation kinetics can lead to the accumulation of misfolded, non-functional intermediates[38]. To evaluate functional output, we quantified the effective protein yield by the bioluminescence intensity of firefly Luciferase and the intrinsic fluorescence of GFP. NCI demonstrated superior predictive power for both proteins (Spearman's r = 0.7 and 0.4) compared to CAI (Spearman's r = 0.6 and 0.2). The particularly strong correlation between NCI and Luciferase functional output likely reflects the reliance of its complex, multi-domain architecture on precise kinetic pacing, whereas the compact β-barrel of GFP is more resilient to kinetic variation. Conversely, the poor performance of tAI (Fig. 6d) suggests that prioritizing raw elongation speed can override the rhythmic pauses essential for nascent chain maturation. By offering robust predictive power independent of target-specific supervision, NCI provides a generalizable framework for the *de novo* design of proteins where experimental data—required by models like Encodon—are unavailable.

Benchmarking 120 diverse proteins (four kingdoms; up to 600 aa), we found that AlphaFold3-derived designs achieved high sequence similarity with those based on experimental

structures (mean 80.9%; *E. coli* 93.4%; human/plant/viral >76%; Fig. 6e). Furthermore, design metrics remained highly consistent between the two structural sources (Spearman's r: CAI=0.936, tAI=0.948, NCI=0.772; Fig. 6f-h). The slightly lower correlation for NCI reflects its sensitivity to fine-grained structural nuances, yet the overall robust agreement demonstrates NatCodon's resilience to the minor variations inherent in AI-predicted models. These results establish AlphaFold3 (AF3) as a high-fidelity alternative, extending NCI-based optimization to the vast majority of the uncharacterized proteome without compromising design performance.

By using NatCodon to generate a library of structure-aware codon sequence candidates and NCI to rank them, we can move beyond simply maximizing protein abundance and instead navigate the complex fitness landscape of translation and folding to achieve predictable optimization of functional protein expression.

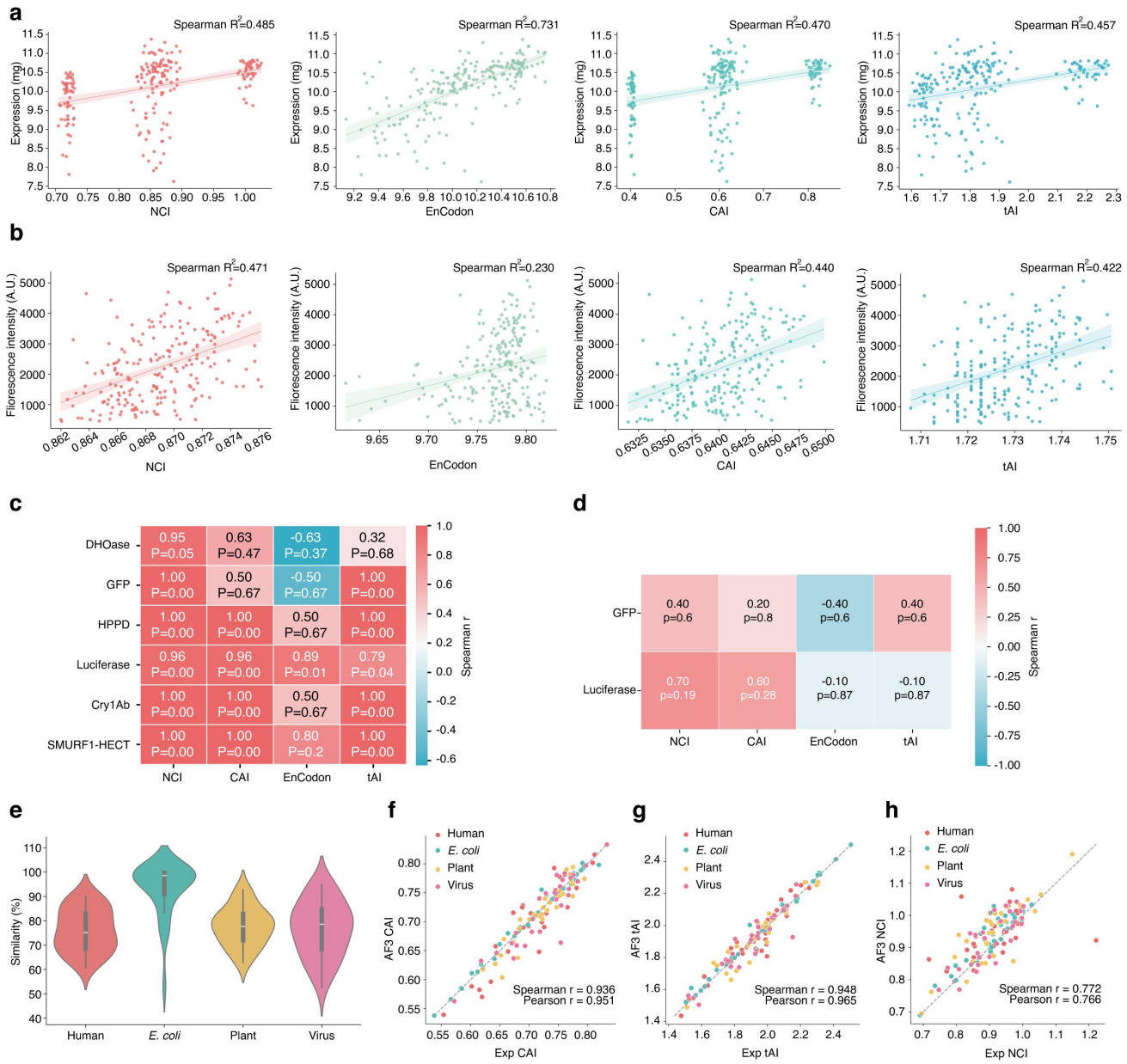


Figure 6. **NatCodon Index (NCI) as a predictor of functional protein output and its compatibility with AI-predicted structures**. **a,b,** Evaluation in large-scale synonymous mutation libraries. Spearman correlations between predicted metrics (NCI, CAI, tAI, and Encodon) and experimentally measured expression or fluorescence are shown for mRFP **(a)** and GFP **(b)** libraries. **c,d,** Association with soluble and functional protein output. **c,** Heatmap showing Spearman

correlation coefficients between different metrics and soluble expression levels for six proteins (including DHOase and $SMURF1_{HECT}$). **d,** Correlation between predicted metrics and functional outputs for Luciferase and GFP. **e,** Sequence similarity between designs based on experimental structures and AF3-predicted structures across 120 proteins from four kingdoms (human, *E. coli*, plant, and viral datasets), shown as violin plots. **f–h,** Concordance between metrics computed using experimental (Exp) and AF3-predicted structures. Scatter plots compare CAI **(f)**, tAI **(g)**, and NCI **(h)** values derived from the two structure sources.

3 Discussion

The Central Dogma in 1958—formalized by Francis Crick's Central Dogma—established the unidirectional trajectory from nucleic acids to functional proteomes as a foundational tenet of molecular biology[45]. However, the reverse process—which would convert protein sequences back to nucleic acid sequences—is precluded by the inherent degeneracy of the genetic code , which gives rise to synonymous codons. While traditional codon optimization has long relied on mimicking host-specific bias to augment heterologous expression, the recent paradigm shift toward de novo protein design has rendered these heuristic approaches increasingly inadequate. The current explosion of AI-generated proteins, which lack evolutionary ancestors or native coding templates, necessitates a generative framework for codon synthesis that transcends simple frequency-based look-up tables. Crucially, existing sequence-centric AI models fail to account for the structure-codon grammar embedded within the nascent polypeptide chain. The decoupling of synonymous codon selection from the spatiotemporal timing of CTF—an inherent flaw in these methods—triggers kinetic mismatches that result in protein misfolding and loss of functional yield. Thus, bridging the gap between deep learning-based sequence generation and the biophysical constraints of protein biogenesis represents a critical frontier in synthetic biology.

In this study, we introduced NatCodon, a protein-structure-informed model that designs codon sequences by integrating the folding requirements of protein with the inherent context of the host to

coordinate translation rhythm with CTF. NatCodon employs a codon-gate mechanism to ensure amino-acid fidelity during codon sequences generation while enabling flexible synonymous codon exploration. In addition, a focus-controlled sampler improves the diversity of generated sequences. To quantify the compatibility between codon sequences and protein folding requirements, we further introduce the NCI, which measures the extent to which a designed codon sequence satisfies the folding constraints of the target protein. Notably, NCI shows a robust positive correlation with functional protein expression. Together, the NatCodon–NCI framework addresses key limitations of traditional codon optimization methods and sequence-based AI models.

Cross-species benchmarking across multi-omics datasets demonstrates that NatCodon preserves native structure-codon grammar beyond simple codon frequency optimization. At the compositional, translational kinetic, and structural levels, NatCodon-generated sequences consistently show higher concordance with native *E. coli* codon sequences than those produced by competing models. Together, these results demonstrate that NatCodon achieves multi-scale host compatibility, simultaneously maintaining compositional balance, preserving translational kinetic landscapes, and safeguarding structure-codon coupling relationships.

We further evaluated NatCodon experimentally through recombinant expression of five structurally diverse proteins and compared its performance with the widely used commercial optimization algorithm GeneOptimizer. NatCodon achieved a 2- to 33-fold increase in soluble

protein yield relative to GeneOptimizer and enabled soluble expression of the otherwise poorly expressed $SMURF1_{HECT}$. These results highlight the ability of NatCodon to mitigate codon-induced misfolding and improve functional protein production.

Protein structural information plays a central role in the success of NatCodon. When structural inputs were removed, soluble expression levels decreased by 4- to 8-fold and, in some cases, were completely abolished—for example, the Cry1Ab codon sequence designed without structural signals produced no detectable soluble protein. Explainable AI analyses further revealed that structural features contribute nearly as strongly as amino-acid sequence information to NatCodon's predictions. The model preferentially assigns high-frequency codons within stable secondary-structure regions while enriching lower-frequency codons in loop regions, consistent with the hypothesis that local translation slowdown facilitates proper folding of emerging polypeptides. Site-directed mutagenesis experiments further confirmed that these structure-dependent codon sequence design principles can significantly influence protein expression outcomes.

To quantitatively evaluate the compatibility between codon sequences and CTF requirements, we introduced the NCI. Across two public datasets, NCI exhibits stronger correlations with protein expression levels than traditional codon usage metrics such as CAI and tAI, and outperforms the AI-based predictor Encodon in predicting expression of GFP variants. In synonymous mutation experiments across six structurally diverse proteins, NCI consistently showed stronger and more

stable correlations with protein expression. It is important to note that Encodon was not originally designed as a general predictor of expression levels. When extensive synonymous mutation datasets are available for a specific target protein, Encodon can provide highly accurate predictions. In contrast, NCI offers broader applicability across proteins lacking such training data.

Beyond improving heterologous protein expression, the NatCodon–NCI framework may introduce new possibilities for metabolic engineering and synthetic pathway optimization. In complex metabolic networks, the objective is often not maximal expression of a single enzyme but rather the maintenance of balanced metabolic flux across multiple pathway steps. By combining NatCodon's diverse codon sequence sampling with NCI-based evaluation, it may become possible to fine-tune the effective expression rates of pathway enzymes so that they match downstream catalytic demands. In this context, decoding the underlying structure-codon grammar may help transform metabolic engineering from an empirical trial-and-error practice into a more predictive, structure-guided discipline.

Despite these advances, several limitations remain. First, the current NatCodon model is primarily trained on *E. coli*, whose proteome likely contains rich signals linking protein structure and codon usage. In future work, extending this framework to additional host systems, including yeast and mammalian expression platforms, will require integrating cross-species data and may benefit from mixture-of-experts architectures. Such extensions could broaden the applicability of NatCodon

to emerging areas such as mRNA vaccines and eukaryotic protein expression systems. Second, the experimental validation presented here is limited to five proteins. Although these proteins span diverse structural topologies and functional classes, larger validation datasets will be required to comprehensively assess generalizability. Third, while explainable AI analyses and mutagenesis experiments support the mechanistic basis of NatCodon's design principles, future studies incorporating ribosome profiling could directly test the predicted effects of codon design on translation dynamics. Finally, although NatCodon implicitly captures several classical determinants of codon optimization—such as mRNA structural features—through host genomic context, these signals are not currently represented as explicit inputs and may warrant further investigation.

Taken together, NatCodon provides a structure-aware framework for codon sequence design that integrates protein folding constraints with host translational context. By preserving structure-codon grammar embedded in natural genes, the NatCodon–NCI framework offers a robust strategy for designing codon sequences for both heterologous proteins and *de novo* AI-designed proteins, and opens new avenues for predictive engineering of protein expression and metabolic systems.

Reference

1. Baneyx, F. & Mujacic, M. Recombinant protein folding and misfolding in Escherichia coli. *Nature Biotechnology* **22**, 1399–1408 (2004).

2. Gąciarz, A. *et al.* Efficient soluble expression of disulfide bonded proteins in the cytoplasm of Escherichia coli in fed-batch fermentations on chemically defined minimal media. *Microbial Cell Factories* **16**, 108 (2017).

3. Hoang, T. A., Nakano, H. & Ojima-Kato, T. A strategy for scalable antibody production: the combination of *Escherichia coli* expression, translation-enhancing peptide and automated refolding system. *Bioscience, Biotechnology, and Biochemistry* **90**, 429–437 (2026).

4. Lobstein, J. *et al.* SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. *Microbial Cell Factories* **11**, 753 (2012).

5. Kapust, R. B. *et al.* Tobacco etch virus protease: mechanism of autolysis and rational design of stable mutants with wild-type catalytic proficiency. *Protein Engineering, Design and Selection* **14**, 993–1000 (2001).

6. Meile, L. *et al.* A high-throughput heterologous expression platform for plant synthetic biology based on Arabidopsis suspension cells. *Journal of Experimental Botany* eraf545 (2025) doi:10.1093/jxb/eraf545.

7. Wilson, R. H., Thieulin-Pardo, G., Hartl, F. & Hayer-Hartl, M. Improved recombinant expression and purification of functional plant Rubisco. *FEBS Letters* **593**, 611–621 (2019).

8. Desai, P. N., Shrivastava, N. & Padh, H. Production of heterologous proteins in plants: Strategies for optimal expression. *Biotechnology Advances* **28**, 427–435 (2010).

9. Rasala, B. A., Muto, M., Sullivan, J. & Mayfield, S. P. Improved heterologous protein expression in the chloroplast of *Chlamydomonas reinhardtii* through promoter and 5′ untranslated region optimization. *Plant Biotechnology Journal* **9**, 674–683 (2011).

10. Ill, C. R. & Chiou, H. C. Gene Therapy Progress and Prospects: Recent progress in transgene and RNAi expression cassettes. *Gene Therapy* **12**, 795–802 (2005).

11. Xie, Q. *et al.* Improved gene therapy for spinal muscular atrophy in mice using codon-optimized hSMN1 transgene and hSMN1 gene-derived promotor. *EMBO Molecular Medicine* **16**, 945–965 (2024).

12. Pardi, N., Hogan, M. J., Porter, F. W. & Weissman, D. mRNA vaccines — a new era in vaccinology. *Nature Reviews Drug Discovery* **17**, 261–279 (2018).

13. Fredens, J. *et al.* Total synthesis of Escherichia coli with a recoded genome. *Nature* **569**, 514–518 (2019).

14. Gibson, D. G. *et al.* Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science* **329**, 52–56 (2010).

15. Gustafsson, C., Govindarajan, S. & Minshull, J. Codon bias and heterologous protein expression. *Trends in Biotechnology* **22**, 346–353 (2004).

16. Lu, C., Guo, L., Fang, B., Shi, J. & Zhou, M. DNA Sequence Changes Resulting from Codon Optimization Affect Gene Expression in Pichia pastoris by Altering Chromatin Accessibility. *JoF* **11**, 282 (2025).

17. Vervoort, E. B. Optimizing heterologous expression in Dictyostelium: importance of 5’ codon adaptation. *Nucleic Acids Research* **28**, 2069–2074 (2000).

18. Bennett, N. R. *et al.* Improving de novo protein binder design with deep learning. *Nature Communications* **14**, 2625 (2023).

19. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

20. CRICK, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).

21. Koonin, E. V. Does the central dogma still stand? *Biol Direct* **7**, 27 (2012).

22. Sharp, P. M. & Li, W.-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucl Acids Res* **15**, 1281–1295 (1987).

23. Lyu, X., Yang, Q., Zhao, F. & Liu, Y. Codon usage and protein length-dependent feedback from translation elongation regulates translation initiation and elongation speed. *Nucleic Acids Research* **49**, 9404–9423 (2021).

24. Zhang, S. *et al.* Alter codon bias of the P. pastoris genome to overcome a bottleneck in codon optimization strategy development and improve protein expression. *Microbiological Research* **282**, 127629 (2024).

25. Reis, M. D. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Research* **32**, 5036–5044 (2004).

26. Fu, H. *et al.* Codon optimization with deep learning to enhance protein expression. *Sci Rep* **10**, 17617 (2020).

27. Jain, R., Jain, A., Mauro, E., LeShane, K. & Densmore, D. ICOR: improving codon optimization with recurrent neural networks. *BMC Bioinformatics* **24**, 132 (2023).

28. Darabi, S. *et al.* Learning the Language of Codon Translation with CodonFM.

29. Fallahpour, A., Gureghian, V., Filion, G. J., Lindner, A. B. & Pandi, A. CodonTransformer: a multispecies codon optimizer using context-aware neural networks. *Nat Commun* **16**, 3205 (2025).

30. Zheng, W. *et al.* Deep-learning-based single-domain and multidomain protein structure prediction with D-I-TASSER. *Nat Biotechnol* https://doi.org/10.1038/s41587-025-02654-4 (2025) doi:10.1038/s41587-025-02654-4.

31. Zhang, H. *et al.* Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature* **621**, 396–403 (2023).

32. Ingolia, N. Ribosome profiling analysis of in vivo translation. in *Regulatory nascent polypeptides* (ed. Ito, K.) 119–133 (Springer Japan, Tokyo, 2014). doi:10.1007/978-4-431-55052-5_7.

33. Wu, C. C.-C., Zinshteyn, B., Wehner, K. A. & Green, R. High-Resolution Ribosome Profiling Defines Discrete Ribosome Elongation States and Translational Regulation during Cellular Stress. *Molecular Cell* **73**, 959-970.e5 (2019).

34. Zhou, M. *et al.* Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* **495**, 111–115 (2013).

35. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**, 218–223 (2009).

36. Komar, A. A., Samatova, E. & Rodnina, M. V. Translation Rates and Protein Folding. *Journal of Molecular Biology* **436**, 168384 (2024).

625 37. Charneski, C. A. & Hurst, L. D. Positively Charged Residues Are the Major Determinants of Ribosomal
626 Velocity. *PLoS Biol* **11**, e1001508 (2013).
627 38. Yu, C.-H. *et al.* Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational
628 Protein Folding. *Molecular Cell* **59**, 744–754 (2015).
629 39. Robinson, P. J. & Bulleid, N. J. Mechanisms of Disulfide Bond Formation in Nascent Polypeptides Entering
630 the Secretory Pathway. *Cells* **9**, 1994 (2020).
631 40. Lee, J., Ju, M., Cho, O. H., Kim, Y. & Nam, K. T. Tyrosine-Rich Peptides as a Platform for Assembly and
632 Material Synthesis. *Advanced Science* **6**, 1801255 (2019).
633 41. Cooley, R. B., Arp, D. J. & Karplus, P. A. Evolutionary Origin of a Secondary Structure: π-Helices as Cryptic
634 but Widespread Insertional Variations of α-Helices That Enhance Protein Functionality. *Journal of Molecular*
635 *Biology* **404**, 232–246 (2010).
636 42. Rudnev, V. R. *et al.* Current Approaches in Supersecondary Structures Investigation. *IJMS* **22**, 11879 (2021).
637 43. Weaver, T. M. The π-helix translates structure into function. *Protein Science* **9**, 201–206 (2000).
638 44. Touvron, H. *et al.* LLaMA: Open and Efficient Foundation Language Models. Preprint at
639 https://doi.org/10.48550/arXiv.2302.13971 (2023).
640 45. CRICK, F. H. On protein synthesis. *Symp Soc Exp Biol* **12**, 138–163 (1958).
641 46. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded
642 and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

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

48. 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.

49. 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).

50. 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).

Method

Dataset construction and structural annotation

To capture the coupling between structural context and translation kinetics, we curated a comprehensive dataset of structure-sequence-gene triplets. We initially retrieved 60,306 bacterial protein crystal structures ($\leq$ 3.0 Å resolution) from the PDB and cross-referenced them with their corresponding CDS from NCBI GenBank. To translate 3D coordinates into a format compatible with generative modeling, we employed the DSSP[46] algorithm to generate a discrete structural sequence; specifically, unstructured regions were designated as 'F', and segments lacking 3D coordinates were labeled as 'X'. To ensure the model learns robust, conserved structure-codon grammars while maintaining high fidelity to the Escherichia coli translational machinery, we implemented a phylogenetically-driven data augmentation strategy. This involved incorporating proteins from evolutionary related bacterial taxa to broaden the structural-sequence space, followed by redundancy reduction via MMseqs2 (90% identity threshold). After rigorous parity filtering to ensure perfect alignment between CDS and PDB-derived primary sequences, we established a final training corpus of 45,126 triplets. Critically, to ensure the host-specific accuracy of our design, the validation and test sets were strictly partitioned to contain only E. coli-specific entries.

Model Architecture

Main Trunk Design. NatCodon is built upon a decoder-only Transformer architecture, drawing

inspiration from the LLaMA framework to facilitate efficient and scalable autoregressive generation. The model consists of four stacked transformer layers, each integrating a multi-head self-attention mechanism and a SwiGLU feed-forward network. This configuration is specifically optimized to capture the non-linear dependencies between local structural motifs and global sequence patterns. To ensure stable gradient flow during training, we employed RMSNorm for pre-normalization and integrated a residual connection strategy across each block.

Integrated Structural-Sequence Encoding. To enable structure-aware codon generation, the model processes interleaved input embeddings derived from both primary amino acid sequences and their corresponding structural tokens. Specifically, For each residue, the amino acid identity and its secondary structure state (e.g., α-helix, β-strand, or custom tokens 'F' and 'X') are projected into a unified high-dimensional representation space. To preserve the spatial and linear contiguity of the protein chain, we implemented RoPE instead of absolute positional encodings. This allows the model to capture long-range structural dependencies and relative positional information with enhanced flexibility, which is critical for modeling the CTF signals that often span across distant sequence regions.

Generative Objectives and Decoding. The model was trained using a composite objective function that balances next-token prediction accuracy with a structure-conditioned constraint. By modeling the conditional probability $P(codon_i|structure_{1:N}, sequence_{1:N}, codon_{1:i-1})$,

NatCodon effectively learns a structure-codon grammar that reconciles the metabolic cost of translation with the thermodynamic requirements of protein folding.

Core Innovations

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[47]. 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.

Focus-controlled Sampling. 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[48,49]. 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 ($\phi * P_{max}$) to zero. A final codon is then drawn via multinomial sampling[50] 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.

Loss Function. Model training was performed using a customized label-smoothing loss function.

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

Where 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

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

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

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

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 $\mathrm{w_{codon}} = \frac{y_{onehot}}{\sum_{c=1}^{C} y_{onehot}}$, which can dynamically adjust according to the distribution of the labels, helping to avoid over representation 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

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

where $\mathrm{P}(\bullet|x_{b.l})$ 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.

NatCodon Index

To quantitatively evaluate how effectively generated sequences align with native structural–translational regulatory patterns, we developed the NatCodon Index (NCI), a holistic composite metric that integrates three orthogonal dimensions of codon usage. The NCI accounts for the structure-aware Codon Adaptation Index (sCAI), which measures codon adaptation relative to local secondary structure contexts, the Structure-Conditioned Codon Divergence (SCCD), which quantifies the statistical deviation of designed codon distributions from the native structure-specific preferences learned by the model, and the tRNA Adaptation Index (tAI), reflecting translational efficiency based on host-specific tRNA pool availability. The NCI is mathematically formulated such that a higher value signifies a higher degree of structure-translation harmony, representing a state where codon selection optimally reconciles the thermodynamic requirements of protein folding with the kinetic constraints of the translational machinery. By encapsulating these multi-faceted biological signals into a single, unified score, the NCI provides a robust proxy for predicting protein expression levels. We hypothesized—and subsequently validated—that sequences with minimized NCI values more closely recapitulate the evolved regulatory logic of natural genes, thereby facilitating superior foldability and yield in the target host.

Implementation and Validation

The model was implemented using the PyTorch framework and trained on an NVIDIA A100 (80GB) GPU for 100 epochs, with convergence consistently observed within the first 20 epochs. To rigorously validate the biological efficacy of NatCodon-designed sequences, we established a diverse protein library encompassing a broad spectrum of expression challenges. This benchmark set included firefly Luciferase as a standard reporter, DHOase as a moderately difficult target, the $SMURF1_{HECT}$—a protein notoriously recalcitrant to soluble expression—and Cry1Ab, which represents an extreme challenge for heterologous systems. All designed coding sequences were synthesized and cloned into the pET-28a(+) expression vector. Recombinant proteins were subsequently expressed in Escherichia coli BL21(DE3) cells under standardized induction conditions to ensure a robust and comparative assessment of protein yield and folding quality across all design variants.

Reference

1. Baneyx, F. & Mujacic, M. Recombinant protein folding and misfolding in Escherichia coli. *Nature Biotechnology* **22**, 1399–1408 (2004).

2. Gąciarz, A. *et al.* Efficient soluble expression of disulfide bonded proteins in the cytoplasm of Escherichia coli in fed-batch fermentations on chemically defined minimal media. *Microbial Cell Factories* **16**, 108 (2017).

3. Hoang, T. A., Nakano, H. & Ojima-Kato, T. A strategy for scalable antibody production: the combination of *Escherichia coli* expression, translation-enhancing peptide and automated refolding system. *Bioscience, Biotechnology, and Biochemistry* **90**, 429–437 (2026).

4. Lobstein, J. *et al.* SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. *Microbial Cell Factories* **11**, 753 (2012).

5. Kapust, R. B. *et al.* Tobacco etch virus protease: mechanism of autolysis and rational design of stable mutants with wild-type catalytic proficiency. *Protein Engineering, Design and Selection* **14**, 993–1000 (2001).

6. Meile, L. *et al.* A high-throughput heterologous expression platform for plant synthetic biology based on Arabidopsis suspension cells. *Journal of Experimental Botany* eraf545 (2025) doi:10.1093/jxb/eraf545.

7. Wilson, R. H., Thieulin-Pardo, G., Hartl, F. & Hayer-Hartl, M. Improved recombinant expression and purification of functional plant Rubisco. *FEBS Letters* **593**, 611–621 (2019).

8. Desai, P. N., Shrivastava, N. & Padh, H. Production of heterologous proteins in plants: Strategies for optimal expression. *Biotechnology Advances* **28**, 427–435 (2010).

9. Rasala, B. A., Muto, M., Sullivan, J. & Mayfield, S. P. Improved heterologous protein expression in the chloroplast of *Chlamydomonas reinhardtii* through promoter and 5′ untranslated region optimization. *Plant Biotechnology Journal* **9**, 674–683 (2011).

10. Ill, C. R. & Chiou, H. C. Gene Therapy Progress and Prospects: Recent progress in transgene and RNAi expression cassettes. *Gene Therapy* **12**, 795–802 (2005).

11. Xie, Q. *et al.* Improved gene therapy for spinal muscular atrophy in mice using codon-optimized hSMN1 transgene and hSMN1 gene-derived promotor. *EMBO Molecular Medicine* **16**, 945–965 (2024).

12. Pardi, N., Hogan, M. J., Porter, F. W. & Weissman, D. mRNA vaccines — a new era in vaccinology. *Nature Reviews Drug Discovery* **17**, 261–279 (2018).

13. Fredens, J. *et al.* Total synthesis of Escherichia coli with a recoded genome. *Nature* **569**, 514–518 (2019).

14. Gibson, D. G. *et al.* Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science* **329**, 52–56 (2010).

15. Gustafsson, C., Govindarajan, S. & Minshull, J. Codon bias and heterologous protein expression. *Trends in Biotechnology* **22**, 346–353 (2004).

16. Lu, C., Guo, L., Fang, B., Shi, J. & Zhou, M. DNA Sequence Changes Resulting from Codon Optimization Affect Gene Expression in Pichia pastoris by Altering Chromatin Accessibility. *JoF* **11**, 282 (2025).

17. Vervoort, E. B. Optimizing heterologous expression in Dictyostelium: importance of 5’ codon adaptation. *Nucleic Acids Research* **28**, 2069–2074 (2000).

18. Bennett, N. R. *et al.* Improving de novo protein binder design with deep learning. *Nature Communications* **14**, 2625 (2023).

19. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

20. CRICK, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).

21. Koonin, E. V. Does the central dogma still stand? *Biol Direct* **7**, 27 (2012).

22. Sharp, P. M. & Li, W.-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucl Acids Res* **15**, 1281–1295 (1987).

23. Lyu, X., Yang, Q., Zhao, F. & Liu, Y. Codon usage and protein length-dependent feedback from translation elongation regulates translation initiation and elongation speed. *Nucleic Acids Research* **49**, 9404–9423 (2021).

24. Zhang, S. *et al.* Alter codon bias of the P. pastoris genome to overcome a bottleneck in codon optimization strategy development and improve protein expression. *Microbiological Research* **282**, 127629 (2024).

25. Reis, M. D. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Research* **32**, 5036–5044 (2004).

26. Fu, H. *et al.* Codon optimization with deep learning to enhance protein expression. *Sci Rep* **10**, 17617 (2020).

27. Jain, R., Jain, A., Mauro, E., LeShane, K. & Densmore, D. ICOR: improving codon optimization with recurrent neural networks. *BMC Bioinformatics* **24**, 132 (2023).

28. Darabi, S. *et al.* Learning the Language of Codon Translation with CodonFM.

29. Fallahpour, A., Gureghian, V., Filion, G. J., Lindner, A. B. & Pandi, A. CodonTransformer: a multispecies codon optimizer using context-aware neural networks. *Nat Commun* **16**, 3205 (2025).

30. Zheng, W. *et al.* Deep-learning-based single-domain and multidomain protein structure prediction with D-I-TASSER. *Nat Biotechnol* https://doi.org/10.1038/s41587-025-02654-4 (2025) doi:10.1038/s41587-025-02654-4.

31. Zhang, H. *et al.* Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature* **621**, 396–403 (2023).

32. Ingolia, N. Ribosome profiling analysis of in vivo translation. in *Regulatory nascent polypeptides* (ed. Ito, K.) 119–133 (Springer Japan, Tokyo, 2014). doi:10.1007/978-4-431-55052-5_7.

33. Wu, C. C.-C., Zinshteyn, B., Wehner, K. A. & Green, R. High-Resolution Ribosome Profiling Defines Discrete Ribosome Elongation States and Translational Regulation during Cellular Stress. *Molecular Cell* **73**, 959-970.e5 (2019).

34. Zhou, M. *et al.* Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* **495**, 111–115 (2013).

35. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**, 218–223 (2009).

36. Komar, A. A., Samatova, E. & Rodnina, M. V. Translation Rates and Protein Folding. *Journal of Molecular Biology* **436**, 168384 (2024).

841 37. Charneski, C. A. & Hurst, L. D. Positively Charged Residues Are the Major Determinants of Ribosomal
842 Velocity. *PLoS Biol* **11**, e1001508 (2013).
843 38. Yu, C.-H. *et al.* Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational
844 Protein Folding. *Molecular Cell* **59**, 744–754 (2015).
845 39. Robinson, P. J. & Bulleid, N. J. Mechanisms of Disulfide Bond Formation in Nascent Polypeptides Entering
846 the Secretory Pathway. *Cells* **9**, 1994 (2020).
847 40. Lee, J., Ju, M., Cho, O. H., Kim, Y. & Nam, K. T. Tyrosine-Rich Peptides as a Platform for Assembly and
848 Material Synthesis. *Advanced Science* **6**, 1801255 (2019).
849 41. Cooley, R. B., Arp, D. J. & Karplus, P. A. Evolutionary Origin of a Secondary Structure: π-Helices as Cryptic
850 but Widespread Insertional Variations of α-Helices That Enhance Protein Functionality. *Journal of Molecular*
851 *Biology* **404**, 232–246 (2010).
852 42. Rudnev, V. R. *et al.* Current Approaches in Supersecondary Structures Investigation. *IJMS* **22**, 11879 (2021).
853 43. Weaver, T. M. The π-helix translates structure into function. *Protein Science* **9**, 201–206 (2000).
854 44. Touvron, H. *et al.* LLaMA: Open and Efficient Foundation Language Models. Preprint at
855 https://doi.org/10.48550/arXiv.2302.13971 (2023).
856 45. CRICK, F. H. On protein synthesis. *Symp Soc Exp Biol* **12**, 138–163 (1958).
857 46. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded
858 and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

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

48. 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.

49. 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).

50. 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).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ExtendedDataFigure2.pdf
- SupplementaryTables13.xlsx
- ExtendedDataFigure1.pdf
- SupplementaryInformation.pdf
- ExtendedDataFigure3.pdf