PACE-3D: A STRUCTURE-AWARE MULTI-DIMENSIONAL EVALUATION FRAMEWORK FOR CODON OPTIMIZATION VIA ALPHAFOLD2-DERIVED METRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Codon optimization is the cornerstone of heterologous protein expression, but current metrics, such as the Codon Adaptation Index (CAI), only focus on translation elongation rates and ignore the complex coupling between translation dynamics and co-translational folding. Therefore, despite theoretically high expression levels, optimized sequences often suffer from misfolding and aggregation. Here, we propose PACE-3D (Protein-Aware Codon Evaluation in 3D), a structure-aware evaluation framework for quantifying the compatibility between mRNA sequences and protein 3D structures. Utilizing the AlphaFold2-derived confidence metrics, pLDDT (predicted Local Distance Difference Test) and PAE (Predicted Aligned Error), we develop six multidimensional metrics to assess local rigidity-velocity matching, domain assembly synchronization, and translational rhythm smoothness, and so on. Analysis of mainstream codon optimization models based on this framework shows that while existing models successfully optimize translation rate and static mRNA structural stability, they severely neglect the dynamic requirements of co-translational folding, leading to ultimate task failure. This work bridges the gap between sequence design and structural biology, providing a universal standard for rational mRNA design in the AlphaFold era.

1 INTRODUCTION

Solubility and efficient expression of heterologous proteins are the cornerstones of synthetic biology, metabolic engineering, and biopharmaceuticals (Baneyx & Mujacic, 2004; Rosano & Ceccarelli, 2014). Because of significant differences in Codon Usage Bias (CUB) among different species, unmodified foreign genes are generally inefficient (Gustafsson et al., 2004). To overcome this obstacle, codon optimization techniques have emerged, which aim to redesign coding sequences through synonymous mutations to adapt them to the host cell's translation mechanism (Welch et al., 2009). Over the past few decades, most mainstream codon optimization strategies have followed the principle of **maximizing translation elongation rates**. The most representative metric is the Codon Adaptation Index (CAI), which predicts expression levels by measuring the codon usage of the target gene and the codon usage of highly expressed host genes (Sharp & Li, 1987). CAI-based algorithms typically replace rare codons with host-preferred high-frequency codons to improve ribosomal translation speed and overall protein yield. In addition, some improved models have incorporated factors such as mRNA secondary structure stability and GC content (Kudla et al., 2009; Boël et al., 2016). These classic computing paradigms have been successful in many simple cases and have become the standard practice in industry and academia.

However, as the demands of biomanufacturing become increasingly complex, this classic paradigm faces serious challenges. Increasing experimental evidence suggests that blindly pursuing high translation rates often leads to the "more haste, less speed" dilemma: although theoretically highly optimized sequences can generate a large number of polypeptide chains, these products often fail to form the correct native conformation and instead aggregate to form biologically inactive inclusion bodies (Agashe et al., 2013; Buhr et al., 2016). This phenomenon is particularly pronounced when expressing structurally complex, long-sequence proteins (Tokuriki & Tawfik, 2009). This profoundly reveals a core conflict: **a successful codon sequence must not only guarantee the quantity of protein (high yield) but also the quality of protein (correct folding)**.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Breakthroughs in modern biophysics have provided new perspectives for understanding this conflict. Protein folding does not begin only after the complete polypeptide chain is synthesised; rather, it is a dynamic process that begins as the nascent polypeptide chain exits the ribosome (Zhang & Ignatova, 2011; Komar, 2009). In this process, the dynamics of translational elongation are crucial. The movement of ribosomes along mRNA is not uniform, but involves programmed acceleration and deceleration (Ingolia et al., 2009). This uneven translation rhythm is believed to be closely related to the structural characteristics and folding requirements of the new chain. For example, at the boundaries of complex structural domains or in regions where long-range contacts need to be formed, a brief pause in the ribosome can provide the necessary time window for the nascent chain to perform the correct conformation search and assembly, thereby avoiding misfolding and aggregation (O'Brien et al., 2014; Kim et al., 2015). Therefore, **the codon sequence not only implicitly encodes amino acid information, but also implicitly encodes the "timetable" that guides protein folding**.

In recent years, deep learning technologies, represented by AlphaFold2 (AF2), have completely transformed the field of protein structure prediction (Jumper et al., 2021). AF2 not only provides high-precision static structural models, but also outputs confidence metrics, namely the predicted Local Distance Difference Test (pLDDT) and the Predicted Aligned Error (PAE), which contain rich information about local rigidity, domain boundaries, and long-range spatial constraints (Tunyasuvunakool et al., 2021; Akdel et al., 2022). These structural insights offer an unprecedented opportunity to re-examine the codon optimization problem from a completely new perspective.

To bridge the gap between sequence design and structural biology, this paper introduces a new structure-aware evaluation framework, **PACE-3D** (Protein-Aware Codon Evaluation in 3D). This framework innovatively utilizes the AF2-derived pLDDT and PAE metrics as proxy indicators of co-translational fold structure requirements to construct six multidimensional quantitative metrics. These metrics aim to systematically evaluate the compatibility between mRNA sequences and protein 3D structures from multiple levels, including local rigidity-velocity matching, domain assembly synchronization, and translational rhythm smoothness, among others. Analysis of the E. coli protein solubility database[1] (eSOL) showed that the correlation between codon frequency and structural indices was significantly stronger in short sequences than in long sequences, which strongly supports the hypothesis that E. coli proteins cannot fold due to the lack of co-translational folding regulation (Agashe et al., 2013). Using the PACE-3D evaluation framework, we conducted a rigorous and critical evaluation of current mainstream codon optimization models. Our analysis aims to reveal the fundamental flaws in existing models when dealing with complex proteins: that is, while they successfully optimize translation speed and static mRNA structural stability, they severely neglect the requirements of dynamic co-translational folding, leading to the failure of the final task. **This work provides a general standard for rational mRNA design in the AlphaFold era and lays the foundation for developing next-generation codon optimization algorithms based on structural dynamics.**

2 Related Work

2.1 Traditional Codon Optimization Strategies

A key challenge in heterologous protein expression is the difference in CUB among different species. To improve expression levels, computer-aided codon optimization has become a standard procedure in synthetic biology and bioengineering.

Historically, the mainstream optimization strategy has been based on the translation efficiency hypothesis. This hypothesis posits that rare codons (corresponding to low-abundance tRNAs) limit the rate of ribosomal translation elongation, thereby reducing the overall protein yield. Therefore, classic indicators such as CAI (Sharp & Li, 1987) have been developed to quantify the consistency between codon usage of target genes and codon usage of highly expressed host genes. CAI-based optimization algorithms typically employ greedy or heuristic methods to replace rare codons with host-preferred high-frequency codons in order to maximize translation speed.

Subsequent studies have shown that considering only codon frequency is insufficient. The secondary structure of mRNA, particularly the stable hairpin structures in the translation initiation region

[1] http://www.tanpaku.org/tp-esol/index.php

(5'UTR and early coding sequences), can significantly hinder ribosome binding and assembly, thereby suppressing translation initiation efficiency (Kudla et al., 2009). Therefore, modern optimization tools, such as GeneOptimizer (Raab et al., 2010) and CodonOpt[2], typically combine CAI maximization with GC content and mRNA Minimum Free Energy (MFE) minimization to ensure transcript stability and accessibility (Kudla et al., 2009).

However, these methods have a fundamental flaw: they model translation as a uniform or smoothly accelerating process, aiming only to maximize yield while largely ignoring protein quality, namely, the correct folding state. Mounting experimental evidence suggests that an excessive pursuit of high translation rates (high CAI) often leads to the aggregation of nascent polypeptide chains before they complete folding, resulting in biologically inactive inclusion bodies. This phenomenon is particularly prevalent when expressing complex multi-domain proteins (Rosano & Ceccarelli, 2014).

2.2 Co-translational Folding Dynamics

The traditional view of protein folding (Anfinsen's dogma) holds that proteins only fold spontaneously into their thermodynamically determined native conformation after the complete polypeptide chain is released from the ribosome. However, modern biophysics has established a new paradigm of "**Co-translational Folding**", revealing that folding begins at the exit of the ribosome tunnel and is a dynamic process that occurs simultaneously with synthesis (Kramer et al., 2009; Cabrita et al., 2010; Rodnina & Wintermeyer, 2016).

In this process, the dynamics of translational elongation are crucial. Ribo-seq data indicate that the movement of ribosomes along mRNA is not uniform but involves programmed pauses (Ingolia et al., 2009). These pauses are not only a sign of inefficient translation, but also have important biological functions. Studies have shown that transient deceleration of the ribosome during the synthesis of critical structural regions (such as domain boundaries, complex secondary structural elements, or sites requiring cofactor binding) can provide the necessary time window for the nascent chain to perform the correct conformational search, thereby preventing the accumulation of misfolded intermediates (O'Brien et al., 2014; Pechmann & Frydman, 2013).

Therefore, an ideal coding sequence should contain structural information to guide folding. However, translating these complex biophysical insights into computational models presents a significant challenge. The main bottleneck lies in the lack of high-throughput data that can quantify protein sequence structure requirements, which makes it difficult for computational biologists to establish a quantitative mapping between translation rate and folding requirements.

3 Methodology

3.1 Data-Driven Analysis and Motivation

The aim of this analysis is to provide quantitative evidence to challenge the pervasive assumption of "speed equals quality" in codon optimization, and to confirm the assumption that long protein folding failures stem from a severe lack of coordination between translational dynamics and structural requirements.

3.1.1 Revealing CAI Limitations:

To evaluate the conventional assumption that maximizing translation elongation rate ensures effective functional expression, we used the eSOL database to examine the relationship between theoretical translation speed (quantified by CAI) and actual protein solubility. Fig.1A shows the solubility distribution of a specific E. coli protein with high theoretical translation efficiency (CAI > 0.7). It is worth noting that, although these proteins exhibit high codon usage patterns according to traditional metrics, a significant portion (over 30%) are still classified as insoluble proteins. This discovery fundamentally challenges the sufficiency of CAI as an important metric of protein quality, showing that simply optimizing codon bias cannot guarantee successful folding and solubilization.

[2]https://www.idtdna.com/CodonOpt

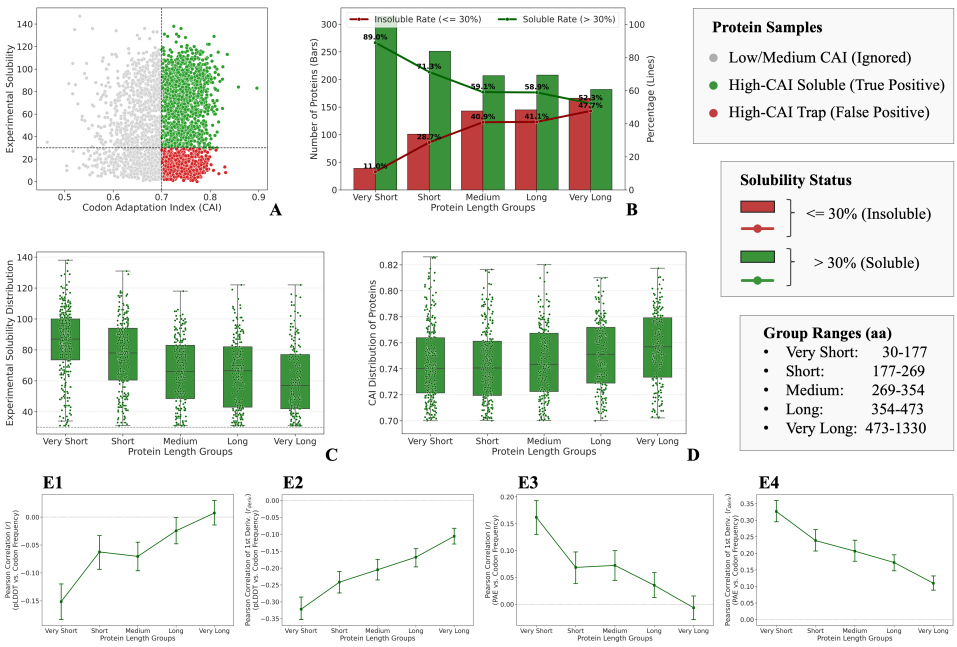


Figure 1: Data-driven analysis and motivation involves data analysis diagrams.

3.1.2 Exposing the Solubility Paradox:

Given that the complexity of protein folding increases significantly with increasing polypeptide chain length, we divided the proteins into five groups based on sequence length to study length-dependent trends. Fig.1B shows the ratio of insoluble to soluble proteins in each group, revealing a clear trend that the proportion of insoluble proteins increases significantly with increasing sequence length. Consistent with this observation, Fig.1C shows that the average solubility score of the shorter sequence group was significantly higher than that of the longer sequence group.

However, a contradictory pattern emerges when examining the codon bias of these protein groups. Fig.1D shows that longer protein groups have a higher average CAI than shorter protein groups. This "**CAI paradox**", where longer proteins exhibit higher theoretical translation speeds yet lower actual solubility, strongly suggests widespread co-translational folding failure in long E. coli proteins. This suggests that high translation speeds, which are not regulated by the specific folding requirements of complex sequences, may be detrimental rather than beneficial to the final protein quality.

3.1.3 Demonstrating Rate-Structure Decoupling:

To verify the hypothesis that this folding failure stems from the "decoupling" of translational dynamics and structural requirements in long proteins, we use AF2-derived metrics as proxies for dynamic structural constraints: pLDDT represents local rigidity requirements, and PAE represents global topological constraints. We calculated the Pearson correlation coefficient between the codon frequency distribution (an indicator of translation speed) and these structural distributions in different length groups.

Fig.1E1 and E3 illustrate the direct correlations between codon frequencies and pLDDT and PAE, respectively. To capture the synchronicity of dynamic changes, Fig.1E2 and E4 show the correlations between the first derivatives of these profiles. A consistent pattern emerged in all four quantitative analyses: the translational dynamics of the short sequence group showed a significantly stronger correlation with structural proxies compared to the long sequence group.

Motivation Summary: In summary, this data-driven analysis provides compelling evidence that while nature has evolved coordinated translation dynamics to meet the structural requirements of short-chain soluble proteins, this necessary coordination is significantly diminished or lost in long-chain, easily aggregated proteins. This observed "rate-structure decoupling" is considered a major

driver of complex protein folding failure. Therefore, a new computational framework is urgently needed to explicitly quantify the dynamic compatibility between mRNA sequences and protein 3D structures.

3.2 THE PACE-3D EVALUATION FRAMEWORK

Based on the data-driven motivation that restoring the coordination between translational dynamics and structural requirements is crucial, we propose the PACE-3D evaluation framework. Fig.4 in Appendix illustrates how PACE-3D is used.

3.2.1 FRAMEWORK PRINCIPLES: DYNAMIC STRUCTURAL CONSTRAINTS:

The core principle of PACE-3D is to reinterpret the static confidence metrics of AF2 as dynamic metrics of co-translational folding requirements. We hypothesize that ribosomes translate not into a predetermined final structure, but into a constantly changing energy landscape defined by the instantaneous folding potential of the nascent chain.

Therefore, we transform the output of AF2 into dynamic constraints:

pLDDT as an indicator of local rigidity requirements: Higher local pLDDT values indicate that residues will form highly ordered rigid structures (e.g., buried cores, stable secondary elements). We hypothesize that synthesizing these regions requires slower translational elongation rates to allow sufficient time for accurate stacking and to prevent kinetic traps leading to incorrect folding states.

PAE as an indicator of global topological and modularity requirements: The PAE matrix encodes spatial adjacency and modularity. Lower off-diagonal PAE values indicate the presence of significant long-range contacts, requiring translational pauses to allow distant sequence elements to meet in 3D space. Conversely, low PAE value regions distributed diagonally define cooperative folding modules requiring coordinated translational rhythms.

Based on these principles, we formulated six multidimensional metrics, divided into two layers: metrics that respond to local rigidity (pLDDT-based) and metrics that respond to global topology and modularity (PAE-based).

3.2.2 LAYER 1: LOCAL RIGIDITY-VELOCITY COORDINATION METRICS BASED ON PLDDT:

These metrics evaluate how well the translation velocity profile (derived from codon usage frequency) adapts to the rigidity requirements of the local structure of the emerging chain.

i. Structure-Translation Correlation (STC)

Rationale: STC provides a macroscopic assessment of the linear relationship between translation speed and structural rigidity throughout the entire protein sequence.

Description: STC is calculated by determining the Pearson correlation coefficient between the smoothed translational velocity profile (V) and the smoothed pLDDT profile (S). Formulas are shown in Appendix and Code.

Interpretation: Based on the "slow-for-rigid" hypothesis, a negative STC value indicates good coordination, meaning that increased structural rigidity leads to a decrease in translational speed. A positive STC value means that it may be possible to pass through a rigid region quickly.

ii. Kinetic Derivative Match (KDM)

Rationale: STC measures the alignment of the original sequence, while KDM assesses dynamic synchronization. It evaluates whether changes in translational velocity (acceleration or deceleration) can predict changes in structural requirements (stiffening or softening).

Description: KDM is the Pearson correlation between the first derivative of the velocity profile (V') and the first derivative of the pLDDT profile (S'), and is downscaled to 0.5. Formulas are shown in Appendix and Code.

Interpretation: Under the "slow-for-rigid" paradigm, a KDM value close to 1.0 indicates high dynamic synergy, where translational deceleration and structural stiffening transformation are appro-

priately synchronized. Conversely, a KDM value close to 0.0 indicates dynamic mismatch, leading to poor adaptation.

iii. Structural Velocity Ratio (SVR)

Rationale: SVR quantifies the strategic velocity allocation of ribosomes between the most demanding (rigid) and least demanding (flexible) regions of a protein.

Description: We divide the sequence into "rigid core" regions (top quartile of pLDDT) and "flexible" regions (bottom quartile of pLDDT). SVR is defined as the ratio of the average translational velocity of the rigid core region to the average translational velocity of the flexible region. Formulas are shown in Appendix and Code.

Interpretation: SVR < 1.0 indicates a protection strategy where the rigid structure employs a deceleration mechanism relative to the flexible loop. SVR > 1.0 indicates a potentially dangerous situation where critical structural components are rapidly passed through.

3.2.3 Layer 2: Global Topology and Modularity Metrics Based on PAE:

These metrics utilize the 2D PAE matrix to evaluate the translational response to complex space environments, module boundaries, and long-range contacts.

iv. Local Structural Entropy-Velocity Correlation (LSEVC)

Rationale: This metric tests the hypothesis that increased environmental complexity leads to slower translation speed. A residue's structural environment is complex ("high entropy") if it must simultaneously satisfy a heterogeneous combination of strong short-range and weak long-range constraints.

Description: We define the "local structural entropy" of residue i as the standard deviation of the i-th row in the PAE matrix. LSEVC is the Pearson correlation coefficient between the structural entropy profile and the translational velocity profile. Formulas are shown in Appendix and Code.

Interpretation: A negative LSEVC value indicates that the ribosome can appropriately slow down when encountering residues with highly complex, heterogeneous spatial constraints.

v. Intra-Module Rate Coherence (IMRC)

Rationale: Proteins typically fold in the form of cooperating subunits or modules. We hypothesize that synthesizing individual rigid modules requires a smooth, coherent translational rhythm to facilitate cooperating folding and thus avoid unstable velocity fluctuations within the unit.

Description: Rigid modules are identified as continuous blocks with consistently low error values along the diagonal direction of the PAE. For each identified module, IMRC is calculated as the coefficient of variation (CV) of the translational velocity within that specific region. Formulas are shown in Appendix and Code.

Interpretation: The lower the IMRC value, the smoother the continuity and transition within the structural module, which is theoretically beneficial for cooperative folding.

vi. Long-Range Constraint Weighted Velocity (LRCWV)

Rationale: This is a direct measurement of topological response. Forming a contact between residues that are far apart in sequence but close in 3D space is a slow and entropic process that requires translational pauses.

Description: For each residue, the "long-range constraint density" is calculated by summing the inverse PAE values of all residues that are in contact with upstream residues at a distance exceeding a preset sequence distance threshold. LRCWV is the Pearson correlation coefficient between this constraint density profile and the translational velocity profile. Formulas are shown in Appendix and Code.

Interpretation: A strong negative LRCWV value is ideal, indicating that the translational velocity decreases significantly with increasing required long-range contact, thus providing the necessary time for topology maturation.

4 EXPERIMENTAL RESULTS AND ANALYSIS

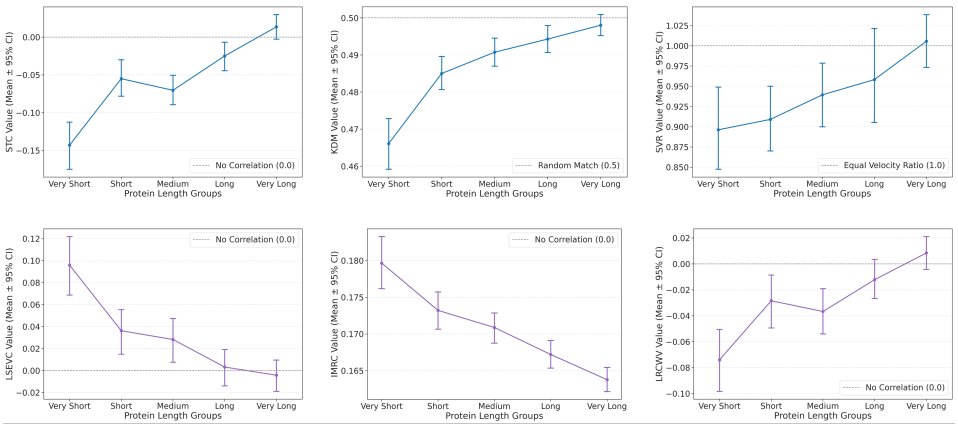


Figure 2: Quantitative analysis of natural E. coli proteins on six metrics within the PACE-3D evaluation framework.

4.1 CO-TRANSLATIONAL FOLDING ANALYSIS OF NATIVE E. COLI PROTEINS

To gain a preliminary understanding of how translation dynamics resonate with structural requirements in nature, we applied the PACE-3D evaluation framework to the E. coli proteome. Using the eSOL database, we categorized proteins into five length groups (from very short to very long) to investigate length-related trends in co-translational folding mechanisms. Fig.2 illustrates the analysis results of native E. coli proteins across the six PACE-3D metrics.

4.1.1 LAYER 1: ANALYSIS OF LOCAL RIGIDITY-VELOCITY COORDINATION

The metrics in this layer evaluate the ribosome's response to immediate structural rigidity requirements. The results reveal that the coordination of "slow-for-rigid" significantly decreases with length.

STC measures global alignment, and the results show that "Very Short" proteins have a strong negative correlation (average of approximately -0.14), suggesting a robust evolutionary strategy of slowing down translation across rigid regions as a whole. However, this coordination gradually weakens as protein length increases, and for "Very Long" proteins, it drops to zero and then becomes slightly positive. SVR also confirms this trend. The short proteins had SVR values significantly lower than 1.0, confirming that they translated the rigid core more slowly than the flexible region. In contrast, SVR increases steadily with length, and for "Very Long" proteins, SVR exceeds 1.0, suggesting that there is maladaptive rapid translation of key rigid core elements in long chains.

KDM reveals different trends by assessing the dynamic synchronicity of the rate of change. KDM values were generally low across all groups (close to 0.5 of the random baseline), but there was a statistically significant increase from "Very Short" to "Very Long" proteins. This suggests that precise dynamic derivative matching may not be the primary mechanism in native E. coli, or that shorter proteins exhibit different dynamic behaviors compared to longer proteins.

4.1.2 LAYER 2: ANALYSIS OF GLOBAL TOPOLOGY AND MODULARITY

The metrics in this layer assess the response to complex spatial constraints and long-range interactions. These metrics also show a clear dependence on protein length.

LRCWV provides strong evidence for length-dependent regulation. The significantly negative LRCWV values of "Very Short" proteins indicate that translation speed is effectively reduced when the required long-range contact density is high. Similar to STC, this favorable coordination is gradually lost with increasing length, approaching zero for "Long" proteins and becoming positive for "Very Long" proteins.

LSEVC and IMRC also showed a clear length-dependent trend. LSEVC was positively correlated with "Very Short" proteins, but gradually turned negatively correlated with long proteins. Similarly, IMRC decreases with increasing length, suggesting that shorter proteins have higher translational variability within structural modules compared to longer proteins.

In summary, the analysis of native E. coli proteins revealed significant length-dependent differences in the regulation of co-translational folding. Short native proteins exhibit strong evidence of evolutionary coordination, particularly global deceleration in rigid regions and long-range constraints (negative STC and LRCWV, low SVR). This coordination is gradually lost in longer proteins, which explains the mechanism by which proteins are more likely to aggregate and demonstrates the necessity of structure-aware optimization.

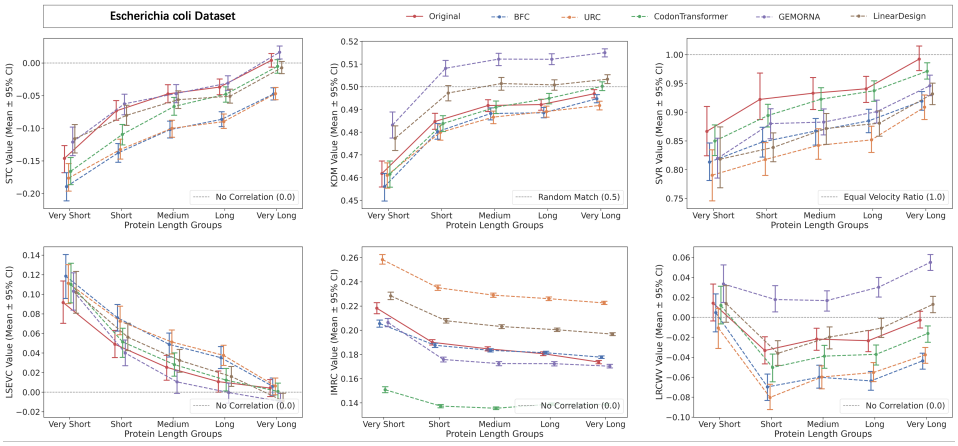


Figure 3: Quantitative analysis of mainstream codon optimization models on the E. coli dataset within the PACE-3D evaluation framework.

4.2 Quantitative Assessment of Mainstream Optimization Models

4.2.1 Experimental Setup:

To provide a comprehensive evaluation of current codon optimization strategies through the lens of co-translational folding, we applied the PACE-3D framework to a diverse set of eight organisms across the tree of life: ***Arabidopsis thaliana***, ***Caenorhabditis elegans***, ***Danio rerio***, ***Drosophila melanogaster***, ***Escherichia coli***, ***Homo sapiens***, ***Mus musculus***, and ***Saccharomyces cerevisiae***. For each organism, we retrieved the proteome-wide predicted structures from the AlphaFold Protein Structure Database[3]. We then generated optimized coding sequences for each protein using five representative mainstream models: ***BFC*** (Background Frequency Choice), ***URC*** (Uniform Random Choice), ***CodonTransformer*** (Fallahpour et al., 2025), ***GEMORNA*** (Zhang et al., 2025), and ***LinearDesign*** (Zhang et al., 2023). The native "***Original***" sequences served as a biologically evolved baseline. Consistent with our previous analysis, proteins within each dataset were stratified into five length groups (Very Short to Very Long) to unravel length-dependent performance characteristics. The performance of each model was quantified using the six PACE-3D metrics, with results presented in Fig.3, as well as Fig.1 through 7 in the appendix.

4.2.2 Detailed Analysis of the Escherichia coli Dataset:

Fig.3 depicts the PACE-3D metric scores for the original and optimized sequences across protein length groups.

In the pLDDT-based metrics (Layer 1): For the STC metric, BFC and URC demonstrate a clear ability to enhance the negative correlation characteristic of co-translational folding across various protein length groups. This suggests that despite the simplicity of their strategies, BFC and URC align with the kinetic requirements of co-translational folding. In contrast, GEMORNA and LinearDesign

[3]https://ftp.ebi.ac.uk/pub/databases/alphafold/v6/

appear to merely capture codon frequency patterns similar to the native mRNA sequences, showing no significant improvement. For the KDM metric, GEMORNA exhibits a positive dynamic translational "rhythm"; however, this trend is opposite to that observed in the original mRNA sequences. Similarly, LinearDesign also deviates from the patterns established by the original mRNA sequences. For the SVR metric, all models and strategies display an intensified trend, indicating faster translation rates in "flexible regions" compared to "rigid core regions".

Overall, considering the overall trends of STC, KDM, and SVR, the models did not achieve improvements on long proteins.

In the PAE-based metrics (Layer 2): For the LSEVC metric, all models demonstrate the ability to "sense" the complexity of the spatial environment surrounding the nascent residue in short proteins and adjust translation speeds accordingly; however, this capability is universally lost in long proteins. The performance of BFC and URC is particularly pronounced, suggesting that simple strategies are sufficient to cope with the spatial complexity of short proteins, which conversely implies that short proteins possess relatively simple spatial structures. For the IMRC metric, URC maintains stable and uniform ribosomal translation speeds within individual, independent rigid structural modules, whereas CodonTransformer exhibits the exact opposite trend. For the LRCWV metric, BFC and URC captured a stronger negative correlation between residues and long-range spatial constraints, while GEMORNA deviated from the patterns observed in original mRNA sequences.

4.2.3 Overview of Other Organism Datasets:

Extending the analysis to seven other eukaryotic and prokaryotic datasets revealed both commonalities and species-specific trends. The recurring pattern observed in simpler organisms, such as Saccharomyces cerevisiae and Caenorhabditis elegans, is similar to that in Escherichia coli: naturally occurring "Very Short" proteins exhibit strong "slow-for-structure" signals (negative STC, low SVR, negative LRCWV), which decay with increasing length, and optimization models often disrupt these signals by increasing translation speed proxies.

In contrast, the baseline behavior of the original sequences differs for more complex multicellular eukaryotes, such as Homo sapiens, Mus musculus, Drosophila melanogaster, and to some extent Danio rerio and Arabidopsis thaliana. For the original protein, the STC of short sequences is usually close to zero or even positive, and increases with increasing sequence length. In these organisms, the difference between the original and optimized sequences is sometimes not very obvious across all metrics, although models like LinearDesign consistently show a higher SVR trend. CodonTransformer often exhibits different behaviors, such as consistently low IMRC and sometimes exaggerated negative peaks in LRCWV, indicating different optimization landscapes.

These results indicate that while the principle of co-translational folding is universal, its specific encoding in codon usage bias varies with species complexity, and mainstream optimization models often fail to adequately preserve these subtle, length-related signals.

5 Conclusion

This study explores the serious shortcomings of traditional "speed-centric" codon optimization strategies when applied to complex proteins. Through comprehensive data analysis, we revealed the "CAI Paradox" and proved that folding failure usually stems from "rate-structure decoupling", in which translation dynamics cannot adapt to the co-translation folding requirements. To address this problem, we developed the PACE-3D (Protein-Aware Codon Evaluation in 3D) framework, which innovatively transforms static AlphaFold2 structural metrics into dynamic constraints. This framework establishes a multidimensional system for quantitatively assessing the fundamental compatibility between mRNA sequences and 3D structural requirements, such as local rigidity and global topology. Experimental applications have demonstrated that PACE-3D can effectively capture the evolutionary rate-structure coordination signals inherent in native proteins. Crucially, comparative analysis shows that ***current mainstream optimization models often prioritize translation speed over structural requirements, thereby disrupting this natural harmony***. These findings highlight the fundamental limitations of existing tools and demonstrate the necessity of incorporating a structure-aware dimension into codon optimization to ensure the expression of functional proteins.

ETHICS STATEMENT

The authors declare that this study adheres to ethical standards. All data used for quantitative and qualitative analysis were obtained from publicly available online sources. The models used are based on authors' open-source code and are appropriately cited. Additionally, Google Gemini and Google Translate were employed solely for language polishing and translation to enhance the manuscript's clarity.

REFERENCES

Deepa Agashe, N Cecilia Martinez-Gomez, D Allan Drummond, and Christopher J Marx. Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Molecular biology and evolution*, 30(3):549–560, 2013.

Mehmet Akdel, Douglas EV Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L Good, Roman A Laskowski, Gabriele Pozzati, et al. A structural biology community assessment of alphafold2 applications. *Nature Structural & Molecular Biology*, 29(11):1056–1067, 2022.

François Baneyx and Mirna Mujacic. Recombinant protein folding and misfolding in escherichia coli. *Nature biotechnology*, 22(11):1399–1408, 2004.

Grégory Boël, Reka Letso, Helen Neely, W Nicholson Price, Kam-Ho Wong, Min Su, Jon D Luff, Mayank Valecha, John K Everett, Thomas B Acton, et al. Codon influence on protein expression in e. coli correlates with mrna levels. *Nature*, 529(7586):358–363, 2016.

Florian Buhr, Sujata Jha, Michael Thommen, Joerg Mittelstaet, Felicitas Kutz, Harald Schwalbe, Marina V Rodnina, and Anton A Komar. Synonymous codons direct cotranslational folding toward different protein conformations. *Molecular cell*, 61(3):341–351, 2016.

Lisa D Cabrita, Christopher M Dobson, and John Christodoulou. Protein folding on the ribosome. *Current opinion in structural biology*, 20(1):33–45, 2010.

Adibvafa Fallahpour, Vincent Gureghian, Guillaume J Filion, Ariel B Lindner, and Amir Pandi. Codontransformer: a multispecies codon optimizer using context-aware neural networks. *Nature Communications*, 16(1):3205, 2025.

Claes Gustafsson, Sridhar Govindarajan, and Jeremy Minshull. Codon bias and heterologous protein expression. *Trends in biotechnology*, 22(7):346–353, 2004.

Nicholas T Ingolia, Sina Ghaemmaghami, John RS Newman, and Jonathan S Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *science*, 324(5924):218–223, 2009.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Soo Jung Kim, Jae Seok Yoon, Hideki Shishido, Zhongying Yang, LeeAnn A Rooney, Jose M Barral, and William R Skach. Translational tuning optimizes nascent protein folding in cells. *Science*, 348 (6233):444–448, 2015.

Anton A Komar. A pause for thought along the co-translational folding pathway. *Trends in biochemical sciences*, 34(1):16–24, 2009.

Günter Kramer, Daniel Boehringer, Nenad Ban, and Bernd Bukau. The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. *Nature structural & molecular biology*, 16(6):589–597, 2009.

Grzegorz Kudla, Andrew W Murray, David Tollervey, and Joshua B Plotkin. Coding-sequence determinants of gene expression in escherichia coli. *science*, 324(5924):255–258, 2009.

Edward P O'Brien, Michele Vendruscolo, and Christopher M Dobson. Kinetic modelling indicates that fast-translating codons can coordinate cotranslational protein folding by avoiding misfolded intermediates. *Nature communications*, 5(1):2988, 2014.

Sebastian Pechmann and Judith Frydman. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature structural & molecular biology*, 20(2):237–243, 2013.

David Raab, Marcus Graf, Frank Notka, Thomas Schödl, and Ralf Wagner. The geneoptimizer algorithm: using a sliding window approach to cope with the vast sequence space in multiparameter dna sequence optimization. *Systems and synthetic biology*, 4(3):215–225, 2010.

Marina V Rodnina and Wolfgang Wintermeyer. Protein elongation, co-translational folding and targeting. *Journal of molecular biology*, 428(10):2165–2185, 2016.

Germán L Rosano and Eduardo A Ceccarelli. Recombinant protein expression in escherichia coli: advances and challenges. *Frontiers in microbiology*, 5:172, 2014.

Paul M Sharp and Wen-Hsiung Li. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research*, 15(3):1281–1295, 1987.

Nobuhiko Tokuriki and Dan S Tawfik. Stability effects of mutations and protein evolvability. *Current opinion in structural biology*, 19(5):596–604, 2009.

Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021.

Mark Welch, Sridhar Govindarajan, Jon E Ness, Alan Villalobos, Austin Gurney, Jeremy Minshull, and Claes Gustafsson. Design parameters to control synthetic gene expression in escherichia coli. *PloS one*, 4(9):e7002, 2009.

Gong Zhang and Zoya Ignatova. Folding at the birth of the nascent chain: coordinating translation with co-translational folding. *Current opinion in structural biology*, 21(1):25–31, 2011.

He Zhang, Liang Zhang, Ang Lin, Congcong Xu, Ziyu Li, Kaibo Liu, Boxiang Liu, Xiaopin Ma, Fanfan Zhao, Huiling Jiang, et al. Algorithm for optimized mrna design improves stability and immunogenicity. *Nature*, 621(7978):396–403, 2023.

He Zhang, Hailong Liu, Yushan Xu, Haoran Huang, Yiming Liu, Jia Wang, Yan Qin, Haiyan Wang, Lili Ma, Zhiyuan Xun, et al. Deep generative models design mrna sequences with enhanced translational capacity and stability. *Science*, 390(6773):eadr8470, 2025.

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A.1 STC: STRUCTURE-TRANSLATION CORRELATION

The Structure-Translation Correlation (STC) quantifies the global linear relationship between the translation velocity profile along the mRNA sequence and the structural rigidity profile of the nascent protein chain.

Let L denote the length of the protein sequence in amino acid residues. We define two vectors of length L:

- **Translation Velocity Profile**, V: Let $V = [v_1, v_2, \ldots, v_L]$, where v_i represents a proxy for the translation elongation speed at codon position i. Typically, v_i is a smoothed value derived from codon usage frequencies (e.g., codon adaptation weights) averaged over a ribosomal footprint window.
- **Structural Rigidity Profile**, S: Let $S = [s_1, s_2, \ldots, s_L]$, where s_i represents the proxy for immediate structural rigidity demand at residue position i. In the PACE-3D framework, s_i is the smoothed AlphaFold2 pLDDT score corresponding to residue i.

The STC is calculated as the Pearson product-moment correlation coefficient between vectors V and S:

$$STC = \frac{\sum_{i=1}^{L}(v_i - \mu_V)(s_i - \mu_S)}{\sqrt{\sum_{i=1}^{L}(v_i - \mu_V)^2} \cdot \sqrt{\sum_{i=1}^{L}(s_i - \mu_S)^2}} \tag{1}$$

where:

- v_i and s_i are the individual elements of the velocity profile V and rigidity profile S at position i, respectively;
- μ_V is the arithmetic mean of the velocity profile V: $\mu_V = \frac{1}{L}\sum_{i=1}^{L} v_i$;
- μ_S is the arithmetic mean of the rigidity profile S: $\mu_S = \frac{1}{L}\sum_{i=1}^{L} s_i$.

The resulting STC value ranges from -1 to +1. Under the "slow-for-rigid" co-translational folding hypothesis, a negative STC indicates favorable coordination, where regions of high structural rigidity (high s_i) are translated at slower speeds (low v_i).

A.2 KDM: KINETIC DERIVATIVE MATCH

The Kinetic Derivative Match (KDM) assesses the dynamic synchronization between translation kinetics and structural demands by evaluating the alignment of their respective rates of change along the sequence.

Let V be the smoothed translation velocity profile and S be the smoothed structural rigidity profile of length L. We define the first derivative vectors representing the discrete rates of change along the sequence:

- **Velocity Change Profile**, V': A vector representing the local acceleration or deceleration of translation. Its elements v'_i are calculated using discrete differentiation of V (e.g., central finite differences: $v'_i \approx \frac{v_{i+1} - v_{i-1}}{2}$).
- **Rigidity Change Profile**, S': A vector representing the local stiffening or softening transitions of the structure. Its elements s'_i are calculated similarly using discrete differentiation of S.

We first calculate the Pearson correlation coefficient, $r_{V',S'}$, between these derivative vectors:

$$r_{V',S'} = \frac{\sum(v'_i - \mu_{v'})(s'_i - \mu_{s'})}{\sqrt{\sum(v'_i - \mu_{v'})^2} \cdot \sqrt{\sum(s'_i - \mu_{s'})^2}}. \tag{2}$$

Under the "slow-for-rigid" paradigm, adaptive regulation implies negative correlation (deceleration $v' < 0$ meets stiffening $s' > 0$). To normalize the metric to a $[0, 1]$ scale where higher values indicate better synergistic matching, the final KDM is defined as:

$$KDM = \frac{1 - r_{V',S'}}{2}. \tag{3}$$

The resulting KDM value ranges from 0 to 1. A value approaching 1.0 indicates high dynamic synergy, where changes in translation speed are inversely synchronized with changes in structural rigidity requirements. A value approaching 0.0 indicates maladaptive dynamic mismatch.

A.3 SVR: Structural Velocity Ratio

The Structural Velocity Ratio (SVR) quantifies the strategic allocation of ribosomal translation speed between the most structurally demanding (rigid) regions and the least demanding (flexible) regions of the protein sequence. It evaluates whether rigid cores are translated slower relative to flexible loops, in accordance with the "slow-for-rigid" folding hypothesis.

Let $V = [v_1, \ldots, v_L]$ be the smoothed translation velocity profile and $S = [s_1, \ldots, s_L]$ be the smoothed structural rigidity profile for a protein of length L.

We segment the protein sequence based on the distribution of rigidity scores in S:

- **Rigid Core Regions** ($\mathcal{I}_{rigid}$): The set of residue indices corresponding to the top quartile (highest 25%) of pLDDT scores. Let $Q_3(S)$ denote the 75th percentile value of S.
$$\mathcal{I}_{rigid} = \{i \mid s_i \geq Q_3(S)\}. \tag{4}$$
- **Flexible Regions** ($\mathcal{I}_{flex}$): The set of residue indices corresponding to the bottom quartile (lowest 25%) of pLDDT scores. Let $Q_1(S)$ denote the 25th percentile value of S.
$$\mathcal{I}_{flex} = \{i \mid s_i \leq Q_1(S)\}. \tag{5}$$

We then calculate the average translation velocity within these distinct regions:

$$\bar{v}_{rigid} = \frac{1}{|\mathcal{I}_{rigid}|} \sum_{i \in \mathcal{I}_{rigid}} v_i \tag{6}$$

$$\bar{v}_{flex} = \frac{1}{|\mathcal{I}_{flex}|} \sum_{i \in \mathcal{I}_{flex}} v_i \tag{7}$$

where $|\mathcal{I}|$ denotes the number of indices in the set.

The Structural Velocity Ratio is defined as the ratio of these average velocities:

$$SVR = \frac{\bar{v}_{rigid}}{\bar{v}_{flex}}. \tag{8}$$

Under the "slow-for-rigid" paradigm, a favorable SVR should be less than 1.0, indicating that the ribosome slows down, on average, when translating rigid core regions relative to flexible regions. An SVR greater than 1.0 suggests potentially maladaptive fast translation through critical structural elements.

A.4 LSEVC: Local Structural Entropy-Velocity Correlation

The Local Structural Entropy-Velocity Correlation (LSEVC) assesses the translational response to the complexity of the local spatial constraints surrounding each residue. It tests the hypothesis that residues situated in structurally heterogeneous environments, characterized by a mixture of strong (low error) and weak (high error) spatial constraints with the rest of the protein, require slower translation elongation to navigate the complex conformational search space.

Let $V = [v_1, \ldots, v_L]$ be the smoothed translation velocity profile for a protein of length L. Let M be the $L \times L$ Predicted Aligned Error (PAE) matrix obtained from AlphaFold2, where the element $m_{i,j}$ represents the predicted error (in Ångströms) of the relative position between residue i and residue j.

First, we define the Local Structural Entropy profile, $E = [e_1, \ldots, e_L]$. The entropy e_i for a specific residue i quantified as the standard deviation of the expected errors associated with that residue across the entire structure (i.e., the standard deviation of the i-th row of the PAE matrix). This measures the heterogeneity of the spatial constraints acting on residue i.

The calculation for the i-th element of the entropy profile is:

$$e_i = \sqrt{\frac{1}{L}\sum_{j=1}^{L}(m_{i,j} - \mu_{i,\cdot})^2} \tag{9}$$

where $\mu_{i,\cdot}$ is the arithmetic mean of the i-th row of the PAE matrix: $\mu_{i,\cdot} = \frac{1}{L}\sum_{j=1}^{L} m_{i,j}$.

Next, the LSEVC is calculated as the Pearson product-moment correlation coefficient between the translation velocity profile V and the local structural entropy profile E:

$$LSEVC = \frac{\sum_{i=1}^{L}(v_i - \mu_V)(e_i - \mu_E)}{\sqrt{\sum_{i=1}^{L}(v_i - \mu_V)^2} \cdot \sqrt{\sum_{i=1}^{L}(e_i - \mu_E)^2}} \tag{10}$$

where μ_V and μ_E are the means of profiles V and E, respectively.

The LSEVC value ranges from -1 to +1. Based on the co-translational folding hypothesis, a negative LSEVC is favorable, indicating that the ribosome slows down (lower v_i) when encountering residues in highly complex spatial environments (higher e_i).

A.5 IMRC: Intra-Module Rate Coherence

The Intra-Module Rate Coherence (IMRC) evaluates the smoothness of the translation rhythm within identified cooperative folding modules. This metric is based on the hypothesis that synthesizing a single rigid structural module requires a coherent, stable translational pace to facilitate cooperative assembly, avoiding erratic speed fluctuations that could disrupt the folding pathway.

This metric involves a two-step process: first, identifying folding modules using the PAE matrix, and second, calculating the rate coherence within these modules.

First, modules are identified as contiguous segments of residues that exhibit consistently low predicted aligned error (PAE) among themselves. Let M be the $L \times L$ PAE matrix. A segment extending from residue index $start$ to end is defined as a module $\mathcal{M}_k$ if the average PAE value within the corresponding sub-matrix on the diagonal is below a predefined threshold τ_{PAE} (e.g., 5 Å).

$$\mathcal{M}_k = \{i \mid start_k \leq i \leq end_k\} \tag{11}$$

subject to

$$\frac{1}{|\mathcal{M}_k|^2}\sum_{i \in \mathcal{M}_k}\sum_{j \in \mathcal{M}_k} m_{i,j} < \tau_{PAE}. \tag{12}$$

This process yields a set of K non-overlapping modules: $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_K\}$.

Next, for each identified module $\mathcal{M}_k$, we quantify the variability of the translation velocity. Let $V = [v_1, \ldots, v_L]$ be the translation velocity profile. The velocity values corresponding to module $\mathcal{M}_k$ are $V_{\mathcal{M}_k} = \{v_i \mid i \in \mathcal{M}_k\}$.The coherence for a single module is measured by the Coefficient of Variation (CV) of its translation velocities. The CV normalizes the variability by the mean speed, allowing comparison across modules transcribed at different average rates.

$$CV_k = \frac{\sigma(V_{\mathcal{M}_k})}{\mu(V_{\mathcal{M}_k})} \tag{13}$$

where $\sigma(V_{\mathcal{M}_k})$ is the standard deviation and $\mu(V_{\mathcal{M}_k})$ is the arithmetic mean of the velocities within module k.

The final IMRC score for the entire protein is calculated as the weighted average of the CVs of all identified modules, weighted by module length to prioritize larger cooperative units:

$$IMRC = \frac{\sum_{k=1}^{K} |\mathcal{M}_k| \cdot CV_k}{\sum_{k=1}^{K} |\mathcal{M}_k|}. \tag{14}$$

The IMRC value is non-negative. A lower IMRC value indicates higher coherence (less variability) in translation speed within structural modules, which is theoretically favorable for cooperative folding. A higher IMRC suggests erratic translational rhythms within modules.

A.6 LRCWV: Long-Range Constraint Weighted Velocity

The Long-Range Constraint Weighted Velocity (LRCWV) assesses the translational response to the immediate demand for establishing long-range tertiary contacts. It is based on the hypothesis that the formation of contacts between the currently synthesized residue and residues far upstream in the sequence is an entropically costly process requiring a translational pause to allow for topological maturation.

This metric involves calculating a profile representing the density of long-range constraints acting on each residue and then correlating this with the translation velocity profile.

First, let M be the $L \times L$ PAE matrix, where $m_{i,j}$ is the predicted error between residues i and j. We define two parameters to classify significant long-range contacts:

- **Sequence Separation Threshold** (d_{seq}): The minimum number of residues along the sequence required to define a contact as "long-range" (e.g., $d_{seq} \geq 12$).
- **Contact Definition Threshold** ($\tau_{contact}$): The maximum PAE value below which two residues are considered to be in significant contact (e.g., $\tau_{contact} \leq 8\mathring{A}$).

For each residue position i (where $i > d_{seq}$), we calculate the Long-Range Constraint Density, d_i. This is defined as the sum of the inverse PAE values for all significant contacts formed with upstream residues j that satisfy the separation threshold. Using the inverse PAE ($1/m_{i,j}$) weights tighter constraints (lower error) more heavily.

Using the Iverson bracket notation $[P]$, which equals 1 if predicate P is true and 0 otherwise:

$$d_i = \sum_{j=1}^{i-d_{seq}-1} [m_{i,j} < \tau_{contact}] \cdot \frac{1}{m_{i,j}}. \tag{15}$$

For initial residues where $i \leq d_{seq}+1$, $d_i = 0$. This results in the profile vector $D = [d_1, d_2, \ldots, d_L]$.

Next, the LRCWV is defined as the Pearson product-moment correlation coefficient between the smoothed translation velocity profile V and the long-range constraint density profile D:

$$LRCWV = \frac{\sum_{i=1}^{L}(v_i - \mu_V)(d_i - \mu_D)}{\sqrt{\sum_{i=1}^{L}(v_i - \mu_V)^2} \cdot \sqrt{\sum_{i=1}^{L}(d_i - \mu_D)^2}} \tag{16}$$

where μ_V and μ_D are the arithmetic means of profiles V and D, respectively.

The LRCWV value ranges from -1 to +1. Based on the co-translational folding hypothesis, a negative LRCWV is highly favorable, indicating that translation speed decreases significantly ($v_i < \mu_V$) as the density of obligatory long-range contacts increases ($d_i > \mu_D$), thereby providing the necessary time for these complex topological interactions to form.

A.7 Limitations

We have drawn correlation conclusions from E. coli and constructed the PACE-3D evaluation framework. Although this framework has been extensively validated across seven species, whether it can be seamlessly generalized to other systems with more complex structures requires further validation.

A.8 Ethical Statement

The authors declare that this study adheres to ethical standards. All data used for quantitative and qualitative analysis were obtained from publicly available online sources. The models used are based on authors' open-source code and are appropriately cited. Additionally, Google Gemini and Google Translate were employed solely for language polishing and translation to enhance the manuscript's clarity.

A.9 THE FIGURES INVOLVED IN THE PAPER

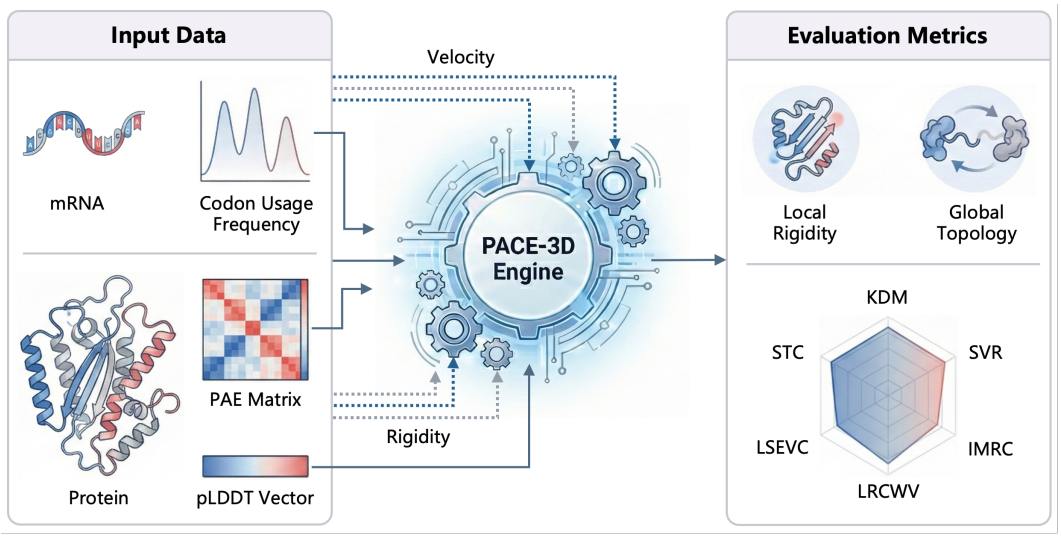


Figure 4: A schematic diagram of the input and output of PACE-3D.

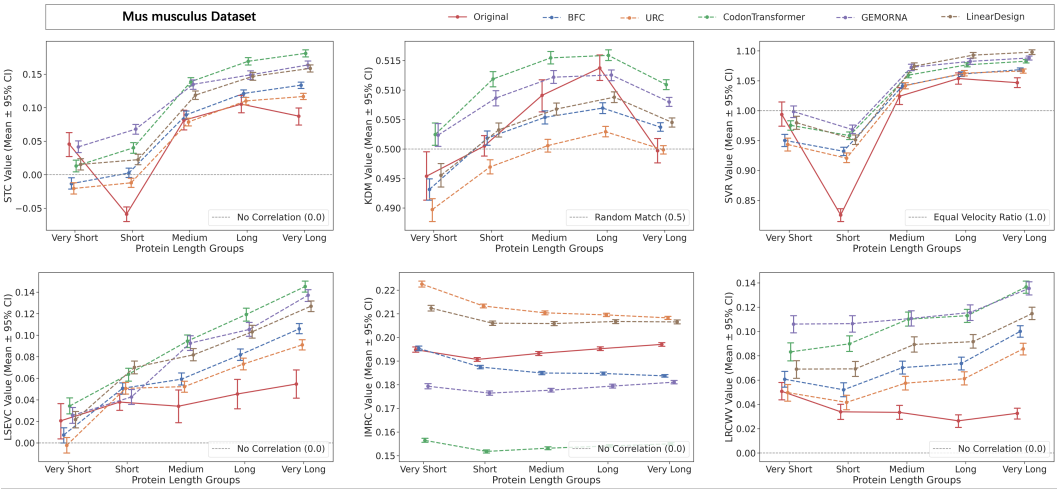


Figure 5: Quantitative analysis of mainstream codon optimization models on the Mus musculus dataset within the PACE-3D evaluation framework.

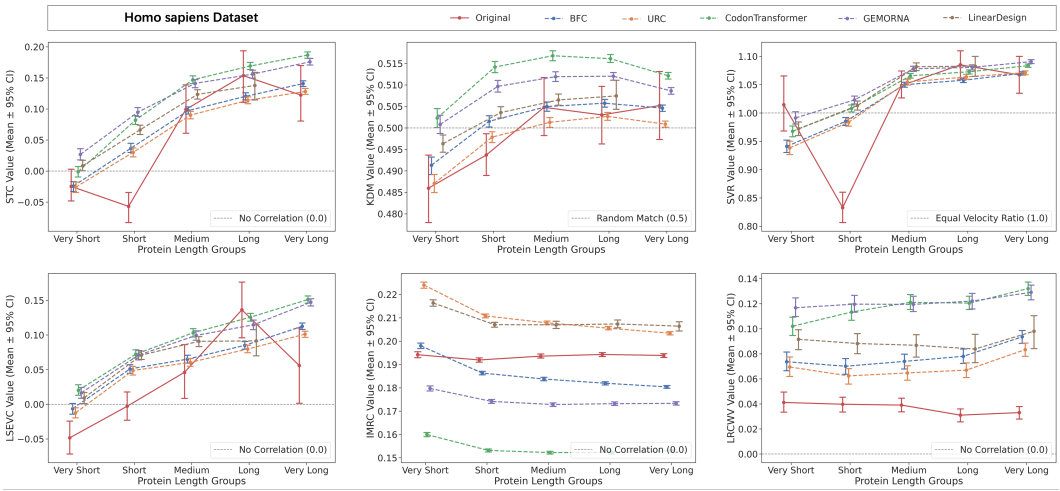


Figure 6: Quantitative analysis of mainstream codon optimization models on the Homo sapiens dataset within the PACE-3D evaluation framework.

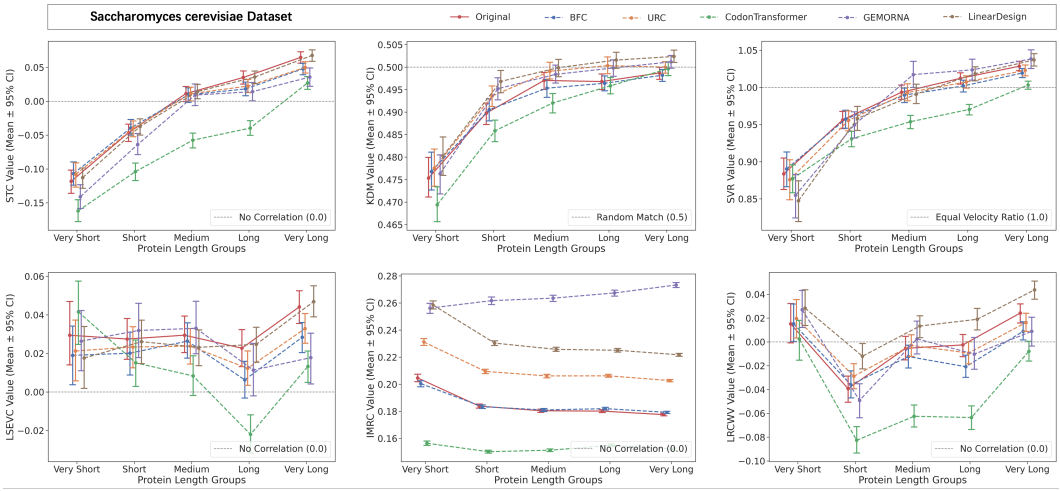


Figure 7: Quantitative analysis of mainstream codon optimization models on the Saccharomyces cerevisiae dataset within the PACE-3D evaluation framework.

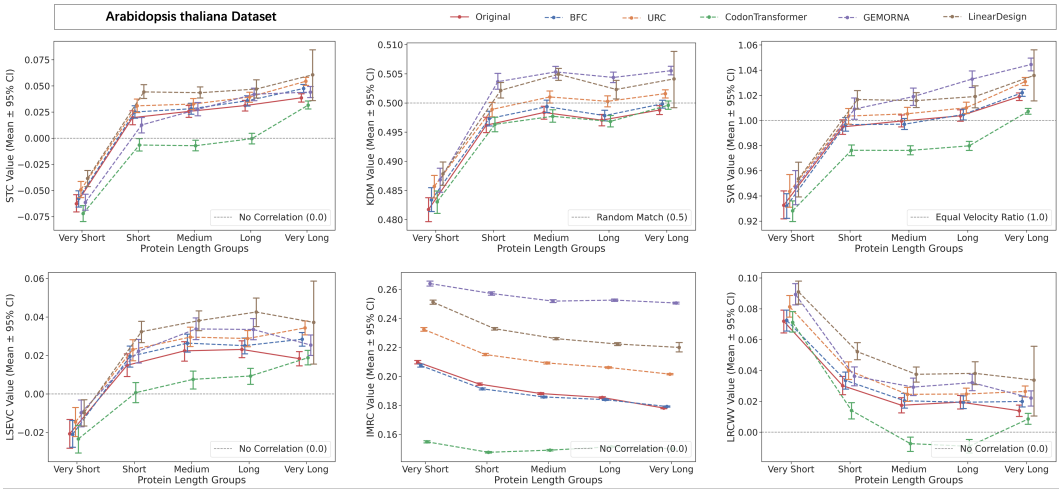


Figure 8: Quantitative analysis of mainstream codon optimization models on the Arabidopsis thaliana dataset within the PACE-3D evaluation framework.

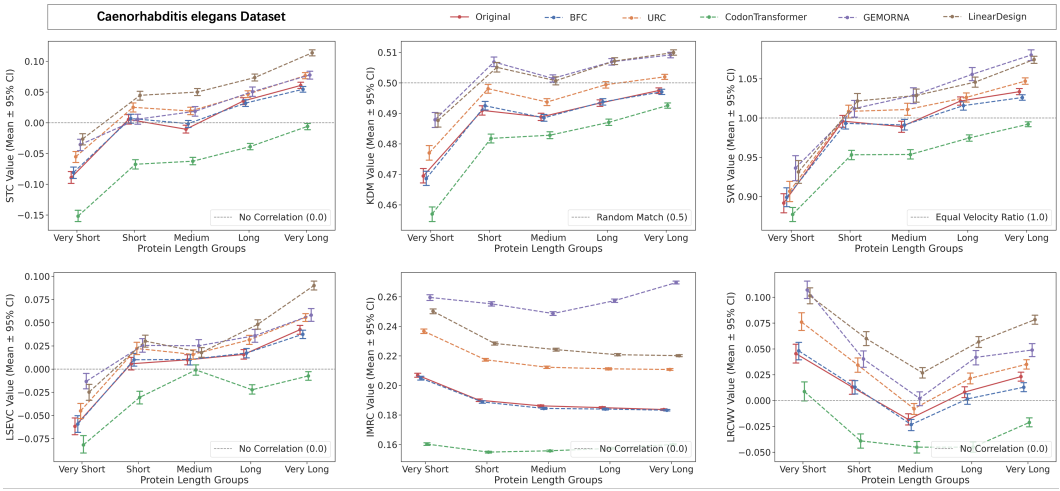


Figure 9: Quantitative analysis of mainstream codon optimization models on the Caenorhabditis elegans dataset within the PACE-3D evaluation framework.

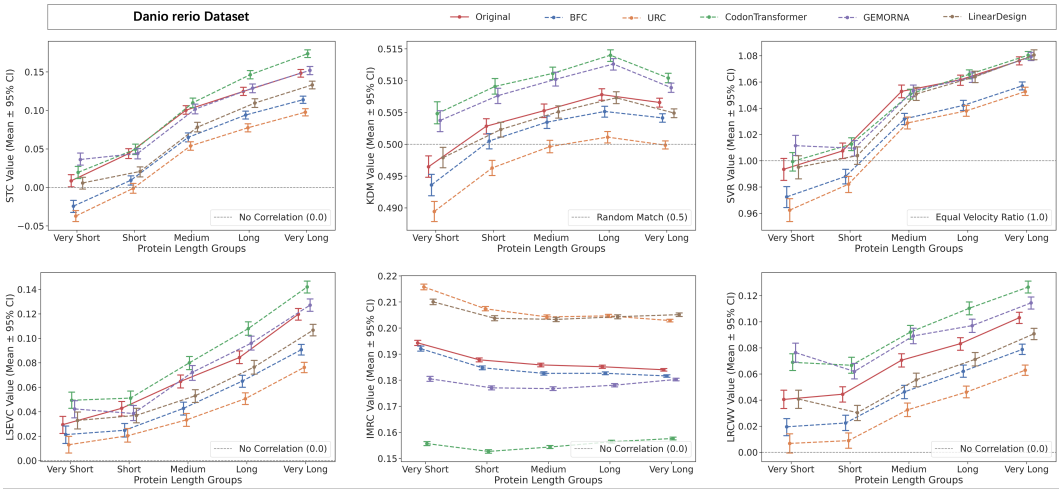


Figure 10: Quantitative analysis of mainstream codon optimization models on the Danio rerio dataset within the PACE-3D evaluation framework.

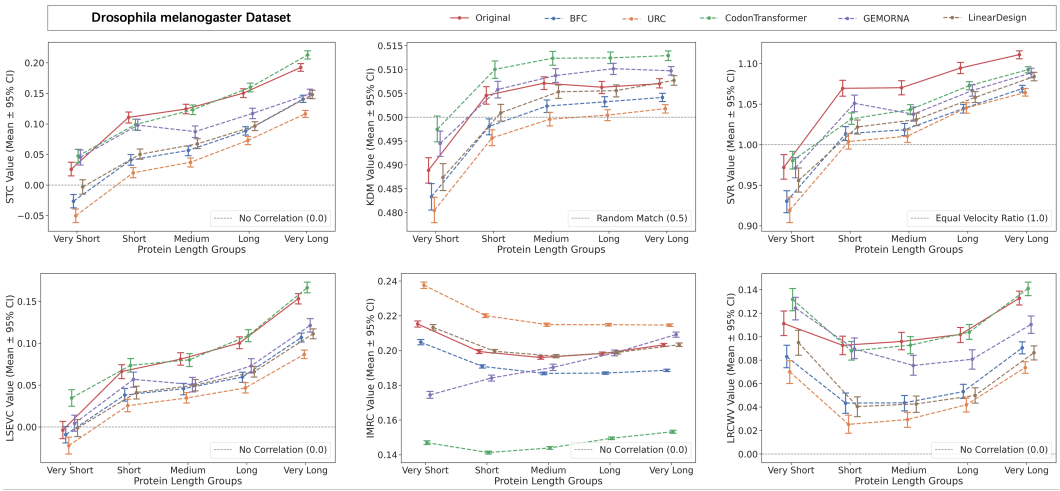


Figure 11: Quantitative analysis of mainstream codon optimization models on the Drosophila melanogaster dataset within the PACE-3D evaluation framework.