

SYNTHETIC BIOLOGY

AI-directed gene fusing prolongs the evolutionary half-life of synthetic gene circuits

Itamar Menuhin-Gruman¹, Matan Arbel-Groissman², Doron Naki²,
Shaked Bergman¹, Tamir Tuller^{1,3*}

Evolutionary instability is a persistent challenge in synthetic biology, often leading to the loss of heterologous gene expression over time. Here, we present STABLES, a gene fusion strategy that links a gene of interest (GOI) to an essential endogenous gene (EG), with a “leaky” stop codon in between. This ensures both selective pressure against deleterious mutations and the high expression of the GOI. By leveraging a machine learning framework, we predict optimal GOI-EG pairs on the basis of bioinformatic and biophysical features, identify linkers likely to minimize protein misfolding, and optimize DNA sequences for stability and expression. Experimental validation in *Saccharomyces cerevisiae* demonstrated substantial improvements in stability and productivity for fluorescent proteins and human proinsulin. The results highlight a scalable, adaptable, and organism-agnostic method to enhance the evolutionary stability of engineered strains, with broad implications for industrial biotechnology and synthetic biology.

INTRODUCTION

Synthetic biology enables the engineering of biological systems for diverse applications, including therapeutic protein production, biosensing, and biomanufacturing (1–14). A key challenge in scaling these systems is maintaining the stability of engineered genes over evolutionary timescales (15–30). Heterologous gene expression often imposes a metabolic burden on host organisms, creating a selective advantage for mutants that reduce or eliminate expression. Over time, this leads to the loss of functionality and impairs the viability of engineered systems for industrial or environmental use (28, 31–36). In addition, this adds regulatory concerns and limits the use of synthetic biology out of the lab, as it leads to a lack of control over the generated sequences (2, 8, 37–39).

Several approaches have been explored to address evolutionary instability of the gene of interest (GOI). Some strategies include generating libraries of complementary parts (40, 41), limiting the user to selected elements, and managing only the problem of repetitive elements—a partial solution. Others entail fine-tuning population dynamics (23, 42–44), requiring designing several strains, tailored solutions, and much experimental tweaking. Many strategies attempt coupling the gene expression of the GOI to the expression of an essential gene and, thus, to host fitness. Some do so by engineering gene overlap (15, 45, 46)—a solution requiring much computational design, which works only for highly specific cases where such overlap is possible. Others require designing biosensors that detect the protein of interest or by-products and activate essential genes on the basis of these biosensors—this solution is highly specific and requires much effort, if possible (47, 48). Yet, another solution suggests using the same promoter for both genes on separate reading frames—mutations to the promoter, which prove deleterious to the GOI, would also prove deleterious to the essential gene, leading to the loss of fitness (22). All these strategies have shown some success but are limited by technical complexity, preventing only specific

mutation types, lack systematic tools, or are constrained to specific settings, genes, or organisms.

Here, we introduce STABLES (stop codon–tunable alternative bifunctional mRNA leading to expression and stability). It is a comprehensive approach to enhancing evolutionary stability through gene fusion. Our strategy involves physically linking the GOI to an endogenous gene (EG) via a shared promoter, on a single open reading frame (ORF), coupled with a “leaky” stop codon to enable differential expression levels (49–52). To optimize this system, we developed a machine learning (ML) tool that predicts the best EG partners for a given GOI, selects linkers likely to minimize protein misfolding using biophysical models of disorder (53–60), and generates codon-optimized DNA sequences for stability and high expression (21, 61–63).

We validated STABLES in *Saccharomyces cerevisiae* by stabilizing the expression of fluorescent proteins and the industrially relevant protein human proinsulin (64). The GOI fused to selected EGs showed greatly enhanced stability and production over successive generations compared to controls. This study provides a scalable, flexible framework for stabilizing synthetic genes, offering a broad range of potential applications in synthetic biology and biomanufacturing.

RESULTS

Overview of the STABLES fusion strategy

Here, we introduce STABLES, a comprehensive approach to enhancing evolutionary stability through gene fusion. It is host and GOI agnostic and robust to many mutations and provides a generic, systematic, simple framework. Our design includes the following components (Fig. 1):

- 1) The GOI to be expressed in the host organism.
- 2) An EG selected for optimal gene expression and mutational stability using an ML model. This model is based on meaningful bioinformatic features and empirical data in *S. cerevisiae*. The GOI and EG are expressed on a shared promoter, on a single ORF, where the GOI's C terminus is fused to the EG's N terminus.
- 3) The linker is selected to minimize disruption to protein folding by comparing disorder profiles of the GOI and EG before and

¹Department of Biomedical Engineering, Tel Aviv University, Tel Aviv 6997801, Israel. ²Shmunis School of Biomedicine and Cancer Research, The George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel. ³The Sagol School of Neuroscience, Tel Aviv University, Tel Aviv 6997801, Israel.

*Corresponding author. Email: tamirtul@tauex.tau.ac.il

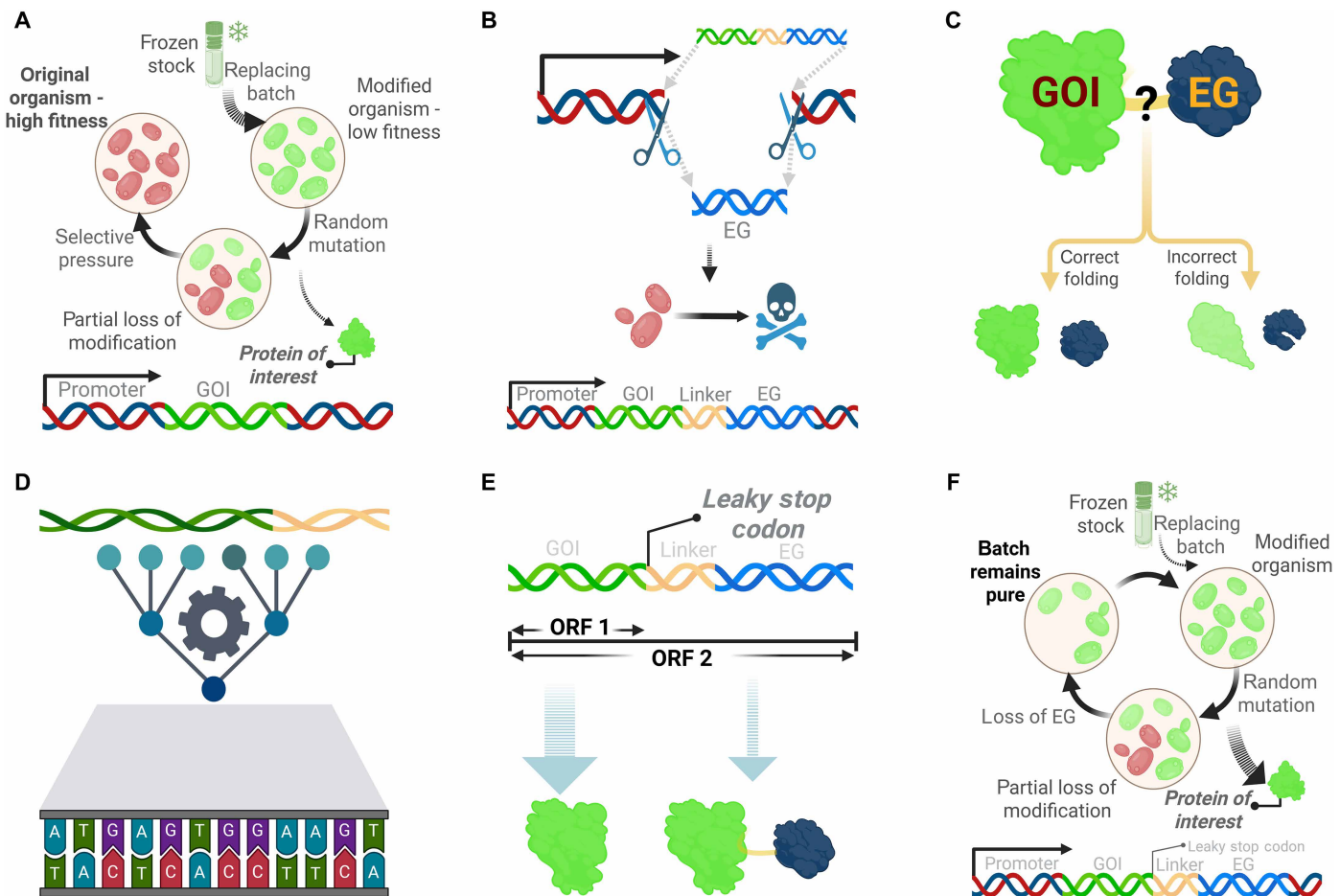


Fig. 1. Components of the STABLES solution. (A) Standard production of a heterologous gene. The heterologous GOI is inserted into the genome. Any mutations that reduce expression or induce misfolding would prove advantageous because of the lower metabolic burden. These mutations proliferate, and a batch must be replaced. (B) Replacement of an EG with a fusion gene. An EG is removed and replaced by a fusion gene composed of the GOI and EG. Mutations leading to the loss of expression or misfolding would be deleterious to the EG, resulting in host death. This limits the spread of many mutations. The EG is selected by an ML model trained on experimental data. (C) Selection of linker. Different linkers may lead to interaction between the fused proteins and misfolding. Using biophysical models and a database of fusion linkers, a linker is selected to minimize structural changes between the fused and unfused states. (D) Sequence optimization. By optimizing the sequence of the GOI and linker, hypermutable sites are avoided, codon usage bias is maximized, and weak mRNA folding is enforced at the start of gene. This further improves stability and expression. (E) A leaky stop codon enables the translation of both the GOI and fusion gene. A leaky stop codon is placed between the GOI and the linker. Because of partial read-through, both the protein of interest and fusion protein are generated. By informed selection of stop codon, large quantities of the protein of interest and just viable quantities of the fusion protein are produced. The GOI's mutational stability is further enhanced, as more mutations would prove deleterious. (F) Production of the heterologous gene, aided by STABLES. The adapted process has much higher mutational stability, reducing the need to replace batches. The optimized cells exhibit much higher expression.

after fusion using biophysical models. A commercial linker yielding a minimal change is chosen (see Methods) (53–60).

4) The fusion gene is optimized for gene expression and the avoidance of mutationally unstable sites (21, 61–63). This includes optimization of the GOI, linker, and depending on use case, the EG as well.

5) A leaky stop codon is placed after the GOI. This is a stop codon with a positive rate of read-through. This leads to the generation of two proteins—either the GOI alone or the fusion protein. The rate of expression for the two proteins can be controlled through the selection of an appropriate rate of read-through (49–52). A codon is selected such that the fusion protein is produced in barely viable quantities for the host's growth while maintaining much higher expression in favor of the GOI's protein alone—as this is the derived, relevant product. By ensuring just barely viable quantities

of the fusion protein, the host's mutational stability is further enhanced, as many more mutations would prove deleterious.

6) The EG in its native form is deleted from the host and replaced by this gene fusion. The host is now dependent on the fusion protein to provide the original EG function. Many mutations that would have reduced production or caused the misfolding of the GOI, whether in the GOI or promoter, would now reduce the production of the fusion protein beneath viable quantities. Because of the host's dependence on this protein, this leads to host lethality, and these mutations do not take hold in the population.

Fusion strategy improves evolutionary stability

To assess the impact of the fusion strategy, we evaluated the stability of green fluorescent protein (GFP) fused to various EGs by examining

10 strains from a previously described library of N-terminally GFP-tagged genes in *S. cerevisiae* (65, 66). Fluorescence was used as a proxy for expression for 15 days.

This experiment was conducted before the creation of the EG selection mechanism—it was designed to give an indication of the need for such a model and its potential impact. For this purpose, meaningful bioinformatic features were generated for all EGs, and they were clustered in many clustering configurations. Ten strains were selected such that they were consistently classified to different clusters and were near the centroids of these clusters. This generated a set of strains that were highly varied between them while being representative of many similar genes. They were compared with a baseline strain of unfused GFP (see Methods).

Fluorescence intensity was used as a proxy for functional GFP levels, following established SWAp-Tag library protocols (65, 66). Given that GFP fluorescence only occurs after proper β barrel folding and chromophore maturation (67–71), it directly reflects the abundance of correctly folded, functional protein. This approach has been validated in large-scale GFP-tagging screens (65, 66, 72, 73) and supported by studies showing that fluorescence changes arise from protein folding and production rather than intrinsic chromophore brightness (74). By contrast, Western blotting detects all GFP-containing species, including misfolded forms, and is prone to technical variability such as antibody binding and transfer efficiency (75–79).

The experiment (Fig. 2A) validated the following conclusions, emphasizing the need for an EG selector:

- 1) Most strains demonstrated decline over the course of the experiment, emphasizing the existence of mutational instability and the need to improve it.
- 2) GOI-EG fusions exhibited slower declines in fluorescence compared to unfused GFP, confirming that the fusion of genes enhances stability.
- 3) Different EGs yielded varying degrees of stability, demonstrating that the stability of a fused gene depends on the EG selected.
- 4) One gene displayed a statistically significant (Student's *t* test, $P \approx 0.047$) advantage over unfused GFP, emphasizing the need for more informed gene selection.

For statistical analysis, see Supplementary 1.

ML predicts optimal EG-GOI combinations

The variability in stability observed across different EGs highlighted the importance of systematic EG selection. To address this, we developed an ML model to predict EG-GOI fusions that maximize expression and stability (Fig. 2B; see Methods). The model was trained on fluorescence data collected from GOI-EG fusion libraries under various conditions in *S. cerevisiae* (see Supplementary 2) (65, 66). As the fluorescence was measured after the variants had time to mutate, this is assumed to capture a combination of both expression and stability.

Using features such as codon usage bias [tRNA adaptation index (80) and codon adaptation index (81)], GC content (82), mRNA folding energy (83, 84), ChimeraARS scores (85), and other meaningful bioinformatic features, the model successfully ranked potential fusion pairs.

In a reasonable use case, a user would generate and use very few genetic designs. The model would recommend one to three EGs, which the user would validate experimentally, and proceed with the design exhibiting best performance. As the models were trained in cross-validation, each EG received a score equal to the

quantile of its expression within the test set (e.g., 1.0 for the EG with the highest expression). Models were evaluated both on their expected performance (median score of EGs recommended by the model among cross-validations) and on their robustness (likelihood of recommending an EG with a low score). For performance measurement with more common and less relevant metrics, see the Supplementary Materials.

On the basis of these evaluations, an ensemble model combining *k*-nearest neighbors (KNN) (86) and XGBoost (XGB) (87) was selected and trained. The KNN model exhibited a high median score, while the XGB model improved the model's robustness. Selecting the best performance among top 3 candidates, the median score was 0.995, and the scores were above 0.98 ($P < 0.05$, Fig. 2C). When selecting only the top performer, the median score was 0.939, and the scores were above 0.92 ($P < 0.001$, Fig. 2D). These results underscore the predictive power of the ML model and its ability to systematically identify highly performing gene fusions, enhancing the efficiency of fusion design. Feature importance was calculated for the features for further insight (Fig. 2E).

To evaluate whether EG selection depends on the GOI, we first assessed universal features of high-performing EGs. The union of the top 50 EGs across GFP, RFP (red fluorescent protein), and human proinsulin exhibited shorter lengths [mean: 962 nucleotides (nt) versus 1354 nt; $P < 1 \times 10^{-8}$], higher codon adaptation [tRNA adaptation index (tAI): 0.568 versus 0.360; $P < 1 \times 10^{-36}$], and slightly higher GC content (43.4% versus 40.3%; $P < 1 \times 10^{-19}$) compared to the yeast genome, indicating that compact, codon-optimized EGs are universally favorable.

We then compared the top 10 EGs for each GOI to assess specificity. The overlap among the highest-ranked EGs was limited: No gene appeared in the top position for more than one GOI, and only GFP and insulin shared two of five top EGs (Jaccard: 0.25), while other pairs showed no top 5 overlap. At larger sets, moderate overlap emerged (e.g., GFP-insulin top 20 overlap: 14 of 20; Jaccard: 0.54), suggesting that certain EGs are broadly strong performers, but the very top predictions remain GOI-specific.

Feature analysis revealed that RFP-optimized EGs were longer and more codon adapted than GFP and insulin EGs (length: $P < 0.001$; tAI: $P < 0.005$), while the GC content did not differ ($P > 0.4$). All top 10 EGs fall within the top 4% of genome-wide tAI scores (quantile: ≥ 0.96), highlighting strong codon optimization across targets. These findings underscore the value of ML-based ranking to prioritize optimal, GOI-specific EGs.

Validation with proinsulin production

To demonstrate the real-world applicability of our approach, we applied it to stabilize human proinsulin expression in yeast, a biotechnologically relevant system. EGs were selected for high performance in both the XGB and KNN models and additional engineering needs (see Methods). Thus, our model identified two EGs—CAF20 and ARC15—as suitable fusion partners for proinsulin.

A 30-day in-lab evolution experiment was conducted, where expression level measurements were taken every 5 days, using the enzyme-linked immunosorbent assay (ELISA) protocol. The strains tested were the original proinsulin (as patented by Novo Nordisk), the proinsulin as optimized by the Evolutionary Stability Optimizer (ESO) (21), and the optimized sequence fused to CAF20 and ARC15 as fusion genes (Fig. 3A). As in the 10-gene experiment, the unfused proinsulin replaced CAN1.

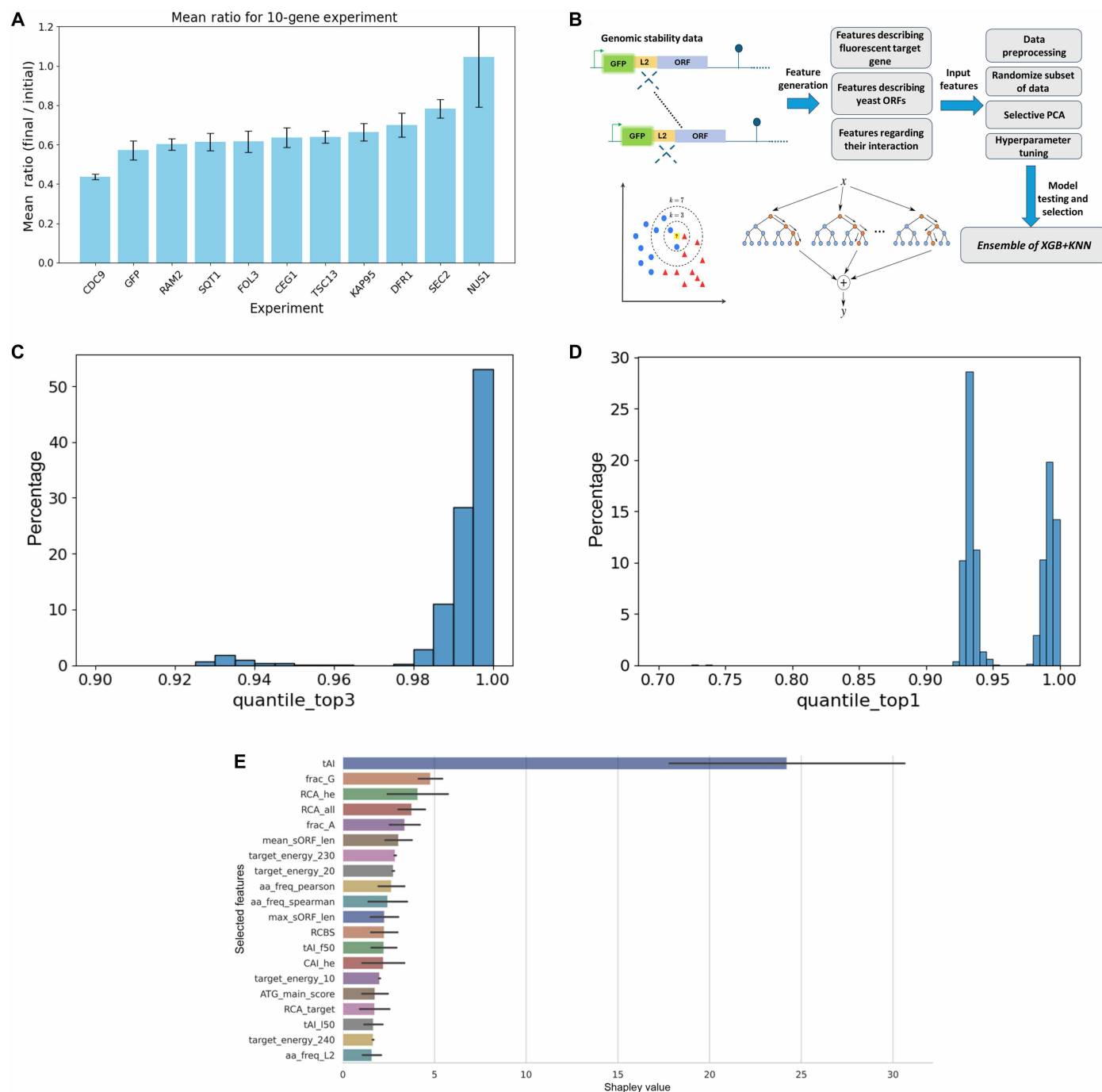


Fig. 2. Computational model for EG selection. (A) Motivation for EG selection model development. Final-to-initial fluorescence ratios for 10 EGs fused to GFP or unfused GFP (second from the left) after 15 days of evolution. Values represent the means from three replicates per condition. Fused constructs retained higher rates of fluorescence than unfused GFP (Student's *t* test, $P \approx 0.048$), indicating greater stability. The observed variability among constructs (Kruskal H, $P \approx 0.035$), emphasizes the importance of rational EG selection. SEC2 showed significantly improved stability over unfused GFP (Student's *t* test, $P \approx 0.047$). We successfully predicted the top performing gene with our model. (B) Gene selection model pipeline. Experimental data (65, 66) covering 6685 EGs and 2 GOIs were used to extract biologically meaningful features, encompassing properties of the EG, GOI, and their interactions. Training and validation splits were used for feature engineering and model tuning. An ensemble of KNN and XGB models was selected on the basis of consistent performance across splits. (C) Performance among top 3 recommendations. In each bootstrap resample, model predictions were converted into quantiles. Model architectures were compared using the highest-performing EG among their top three predictions, reflecting expected performance when testing three fusion constructs. For the selected architecture, the x axis displays the top score among recommendations. The y axis indicates the fraction of bootstrap samples within each performance bin. Predictions were consistently near-optimal (median quantile: 0.995), with a low likelihood ($P \approx 0.048$) of scoring below 0.98. (D) Performance for top recommendation. Similar analysis for the single top recommendation yielded a median quantile of 0.939, with rare scores below 0.92 ($P \approx 0.007$). (E) Distribution of top Shapley values. Averaged across 20 XGB models, tAI emerged as the most predictive feature, followed by GC content, codon usage bias, alternative ORF lengths, mRNA folding energy, and amino acid composition similarity between the EG and the GOI.

The need for a leader secretion sequence in proinsulin production in yeast has been well documented (88–94). However, it was not clear whether it may disrupt the efficacy of our design. We measured the expression at initiation for the original proinsulin and two fusion genes with and without a leader sequence (Fig. 3B). All variants exhibited much higher expression in the presence of a leader sequence, and thus, all further experiments were conducted as such.

The expression patterns approximately followed an exponential decay pattern, where the fused genes displayed a much slower decay rate (Fig. 3C). This enabled us to estimate the cumulative expression of proinsulin over time. These quantities were normalized by the expression estimated for 10 days for the original proinsulin. The gene fusions showed a fivefold increase in total proinsulin yield over the experimental period (Fig. 3D). *ARC15* demonstrated better performance for shorter durations (higher initial expression), while *CAF20* demonstrated better performance for longer durations

(slower decay)—the selection of the optimal gene will depend on the industrial use case.

For the fusion gene with *ARC15*, we measure 98.5 mg/liter at initiation, where for the original sequence, we measure 72.9 mg/liter—a ~35% increase at the initial expression. For the fusion gene with *ARC15*, we measure 68.7 mg/liter after 30 days, meaning that ~70% of expression is maintained. For the original sequence, we measure 4.8 mg/liter, meaning that only ~7% of expression is maintained.

Nanopore sequencing was conducted on the final sequences. It reveals that following the experiment, the fused proinsulin suffered few mutations, while the unfused proinsulin sequence was lost completely (Fig. 4A) (95–98). This has been further validated by Western blotting (Fig. 4B).

We identified several mutations in the *CAF20* and *ARC15* fusion constructs through nanopore sequencing (Fig. 4A and table S4). Notably, all mutations within the proinsulin region were present at similar frequencies both before and after the 10⁵-generation

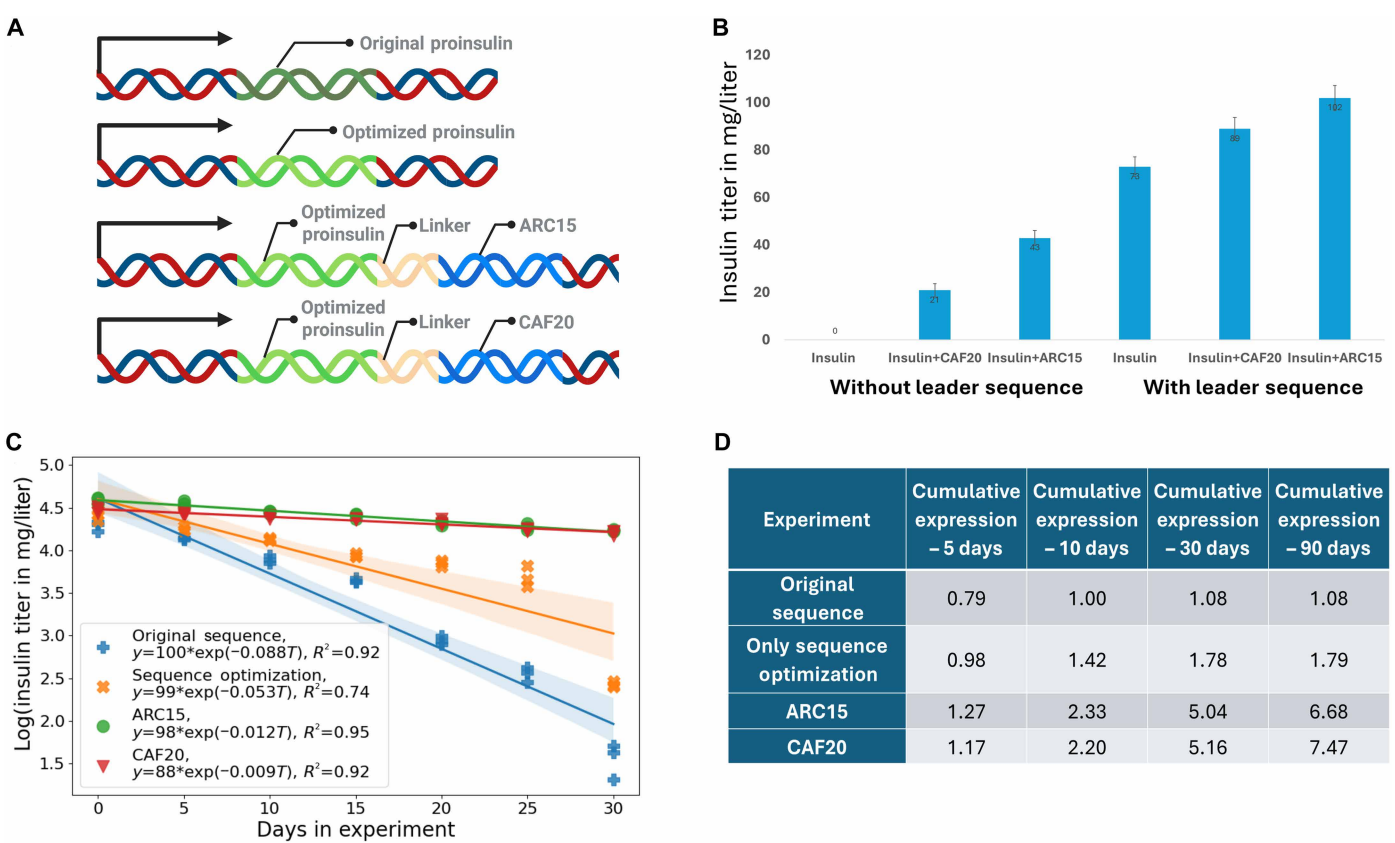


Fig. 3. Demonstration of STABLES: Improving proinsulin expression in *S. cerevisiae*. (A) Design of the evolution experiment. The schematics of the variants in the first proinsulin evolution experiment are displayed for clarity. (B) Gene fusion affects expression at time zero. ELISA measurements of proinsulin expression in six configurations—unfused or fused to *CAF20* or *ARC15* with or without a leader sequence—at day 0. As previously shown, the leader sequence is essential for expressing unfused proinsulin. For fusion constructs, it substantially improves expression, potentially due to increased mRNA stability or improved protein localization in the cell. (C) Proinsulin production over time. ELISA measurements were taken every 5 days ($n = 3$ per time point) for four variants: (i) baseline—Novo Nordisk’s patented sequence, integrated at the *CAN1* locus; (ii) optimized—same structure with proinsulin and linker optimized by ESO (21); (iii and iv) fusions—proinsulin fused to *CAF20* or *ARC15*. Fusion genes were synthesized using the full STABLES pipeline (EG selection, linker choice, leader inclusion, and codon optimization). Expression decay over time fits an exponential model, supported by high R^2 in the log-linear space ($P < 10^{-6}$). The variants exhibit significantly different decay rates [analysis of variance (ANOVA) F -test, $P < 10^{-19}$], with the STABLES-designed constructs showing substantially improved stability. (D) Normalized cumulative proinsulin production. Integrating the fitted expression curves yields the cumulative expression per variant. All values were normalized by the 10-day cumulative expression of the baseline variant. STABLES-derived variants exhibited greatly increased cumulative yields across time points, illustrating improved production through rational EG fusion design.

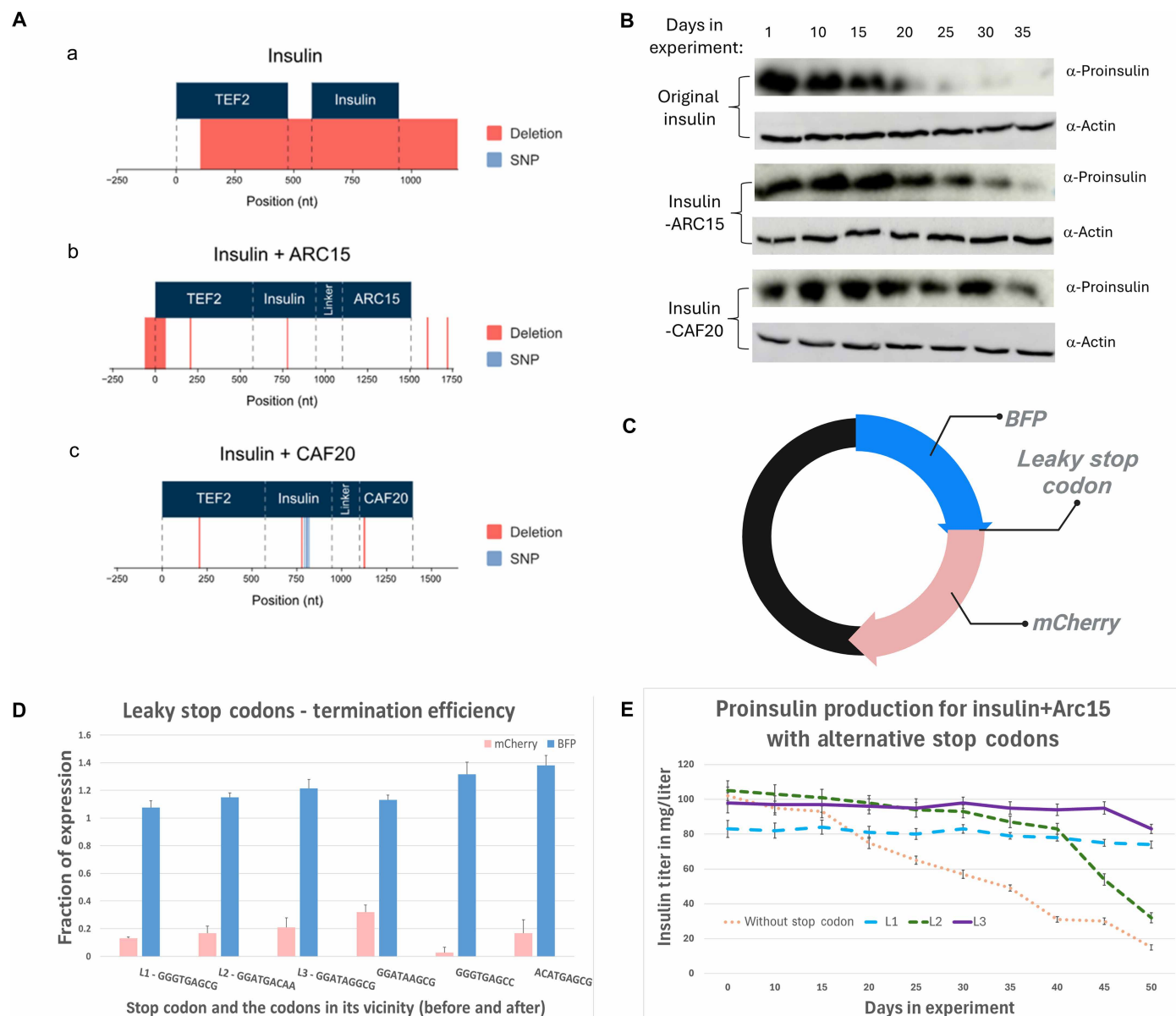


Fig. 4. Experimental study of STABLES results and components. (A) Mutation accumulated in the in-lab evolution experiment. Nanopore sequencing revealed widespread deletions following in-lab evolution: (a) Unfused proinsulin lost most of its promoter and coding sequence; (b) *ARC15* fusion showed large promoter deletions and smaller mutations in both the promoter and proinsulin; (c) *CAF20* fusion retained most of its structure, with only minor mutations across the construct. (B) Western blot validation. Western blotting confirms that *CAF20* and *ARC15* fusions yield prolonged proinsulin expression compared to unfused constructs. Actin measurements were taken for control. (C) Leaky stop codon construct design. Schematic of the construct used to assess translation read-through rates of leaky stop codons using BFP and mCherry fusion. (D) Termination efficiency of leaky stop codons. Various constructs were tested, each linking BFP and mCherry via different leaky stop codons. The fluorescence intensity of each fluorophore (normalized to standalone expression) served as a proxy for read-through efficiency. Three top-performing designs [L1 to L3; see (E)] and three rejected variants (with too high or low read-through) are shown. The BFP signal represents GOI expression, and mCherry represents the C-terminal EG. (E) Proinsulin expression for different leaky stop codons. *ARC15* fusions were built with three different leaky stop codons and without them as the control. ELISA measurements over 50 days demonstrated that all leaky stop codons improved protein stability (Student's *t* test at $t = 40$, $P < 10^{-5}$). Distinct expression profiles among codons (Kruskal H test, $P < 0.003$) emphasize the need for informed selection. All selected codons preserve the last amino acid of proinsulin, use the design principles outlined in (166), and displayed a read-through rate of 0.1 to 0.25 in the previous experiment. Sequences used (stop codon + 3 downstream nt): L1 – TAGGCG; L2 – TGAGCG; L3 – TGACAA.

evolution experiment, suggesting that they were introduced during cloning or represent sequencing artifacts.

This includes a shared 1–base pair deletion (~20% frequency) in a noncoding spacer between the insulin preleader and the B subunit found in both the *CAF20* and *ARC15* constructs; this ostensible deletion of adenine occurs in a homopolymer run of eight adenines, an area highly prone to sequencing error—making this deletion likely to be an artifact.

The *ARC15* construct contained no additional mutations within the proinsulin region. The *CAF20* construct, however, exhibited five silent single-nucleotide polymorphisms in the proinsulin coding region; these mutations were likely introduced during cloning, because they appeared in very high frequencies (>99%), and were stable over time. We also observed a 2-nt deletion in codon 9 of *CAF20*, which increased in frequency from ~40 to ~80% over the course of the experiment and is located within a homopolymer run of five thymines, meaning that this deletion is likely to be at least partially a sequencing error. In addition, the relatively high proinsulin protein levels we observed during the experiment support the conjecture that this is completely or predominantly a sequencing error. This mutation, if it is real and not a sequencing error, introduces a premature stop codon. Despite potential sequencing noise, the observed frequency shift suggests that it may be under positive selection.

Leaky stop codon enables GOI production and overexpression

To enable the generation of both the fusion protein and GOI alone, we used leaky stop codons, which allow partial translational read-through (49–52). The ability to enforce low but nonzero read-through rates is a key factor in improving the stability and industrial viability of our design. If the fusion protein's abundance is reduced to barely viable, then it is highly likely that any mutations affecting the GOI's expression would lead to host lethality, promoting mutational stability. This is in tandem with the enforcement of much higher expression of the isolated GOI, which is the necessary and desired product.

Thus, informed selection of the stop codon is necessary. To enable this informed decision, we conducted an experiment, creating a fusion protein of blue fluorescent protein (BFP) and mCherry, with different leaky stop codons in between. This was expressed on a plasmid (Fig. 4C). The red fluorescence is expressed only in the fusion protein, while the blue fluorescence is expressed in both the fusion protein and BFP alone. Thus, the ratio between the fluorescence measurements can be used to derive the rate of read-through (Fig. 4D).

We selected the three stop codon designs expressing the lowest mCherry fluorescence (to minimize read-through), which was still detectable (to increase cell viability). We conducted another evolution expression, attaching the proinsulin gene to *ARC15*, either without a stop codon or with one of the three selected. This is in accordance with the STABLES design. Using a similar protocol, the expression of proinsulin was measured over 50 days, this time capturing the presence of both the fusion protein and isolated proinsulin. In addition to the secretion of isolated proinsulin (which is a requirement in and of itself), we observed a much higher expression rate and cumulative expression accordingly (Fig. 4E).

For the fusion gene without a stop codon, we measure 102 mg/liter at initiation, where for the best candidate L3, we measure 98 mg/liter—a ~4% decrease at the initial expression. For the best candidate

L3, we measure 83 mg/liter after 50 days, meaning that ~85% of expression is maintained. For the fusion gene without a stop codon, we measure 15 mg/liter, meaning that only ~15% of expression is maintained.

DISCUSSION

Organism-agnostic strategy to tackle evolutionary instability

Previous studies proposed methods for increasing the mutational stability of the GOI. Some of these methods either are dependent on host-specific pathways or require design and implementation of complex parts (e.g., biosensors and overlapping genes), which may not be possible. Others require developing multiple strains and managing population dynamics or provide robustness only to specific mutation types (e.g., repetitive parts and mutations in the promoter). This study introduces STABLES, a simple, robust, generic, host and GOI-agnostic strategy to address evolutionary instability in synthetic biology. By fusing the GOI to an EG on the same ORF under a shared promoter, many deleterious mutations would lead to host lethality, leading to higher GOI mutational stability. By incorporating a leaky stop codon, we enable these gene fusions to generate high expression of the GOI alone while increasing their mutational stability. These, together with informed selection of the linker, sequence optimization, and the use of a leader sequence, are key in the improved stability and performance observed in our experiments.

We find that following all our design principles, we observe a 15% decline in the expression of proinsulin over 50 days. This contrasts with the original design, which declined by 93% over 30 days. This is in addition to a ~30% increase in the initial expression.

Our ML tool enhances the utility of this approach by selecting better GOI-EG pairings on the basis of biologically meaningful features. Translational efficiency metrics, such as tAI, emerged as dominant predictors. A possible explanation for this result may be related to the fact that higher translation efficiency tends to be associated with genes that are highly expressed, have higher mRNA levels (because of higher mRNA stability, among other factors), and are more conserved and thus tend to fold more efficiently (99–103). Other features include GC content, RNA folding energy, alternative shifted ORFs, and amino acid composition that are probably also associated with higher expression, mRNA stability, and robust protein folding.

Protein disorder predictions were used to guide linker selection, increasing the likelihood that the gene fusions maintain native protein folding and functionality (see Methods). Combined with the avoidance of mutational hotspots and sequence optimization for expression and stability, these elements ensure the practical applicability of the approach across a broad range of use cases.

Although our experiments focus on *S. cerevisiae*, the STABLES framework is broadly generalizable to other organisms. First, the general idea of coupling a heterologous GOI to an essential gene is applicable to any living cell. If performed accurately, such a coupling will induce an addiction of the cell to the fused protein, which will prolong the half-life of the GOI. Moreover, its computational components are designed for cross-species applicability: The essential gene selection model uses organism-agnostic, sequence-based features; linker design relies on biophysical and ML models of protein disorder, which are host-independent; and codon optimization adapts

directly to the target organism via the tAI maximization. Leaky stop codons, a key feature of STABLES, have been reported in diverse taxa, including bacteria (104, 105), fungi and animals (49), insects and nematodes (106), and mammals (107, 108). Together, these properties make STABLES a flexible and transferable strategy requiring minimal host-specific adaptation.

Future directions

Using our ML model, we have already seen empirical evidence for good performance in predicting high-performing GOI-EG pairs for different GOIs, as shown in our experimental validation. Its reliance on sequence-based features enables effective application to nonmodel organisms, even in the absence of extensive empirical data. However, expanding the dataset used for training the model remains a valuable avenue for further enhancement. Systematic testing for libraries from more host organisms and a broader range of target genes would improve the model's generalizability and ensure its utility across an even wider variety of synthetic biology applications. Furthermore, in many organisms, epigenetic silencing must be considered when optimizing expression and stability (24, 109–116). While we have designed our model with capabilities to avoid epigenetically silencing motifs, this application should be tested *in vivo*.

While our method has shown much promise, further experimental validation is needed to refine certain aspects. For example, empirical testing of linker selection will help confirm computational predictions and guide future improvements. An in-depth analysis and testing of the leaky stop codons would also enable further optimization.

Broader implications

This study demonstrates the potential to stabilize synthetic genes in both industrial and environmental contexts. In biomanufacturing, the ability to maintain stable GOI expression over extended periods can reduce costs, improve scalability, and simplify regulatory processes. In environmental applications, robust gene fusions could support long-term deployment under dynamic and uncontrolled conditions, enabling breakthroughs in bioremediation and biosensing.

METHODS

Overview of gene fusion design

We developed a multistep workflow to design and optimize GOI-EG fusion genes for enhanced evolutionary stability and expression. This process included the selection of EGs using an ML model, the identification of optimal linkers to minimize misfolding, sequence optimization of the synthetic gene to maximize expression and stability, and addition of a low-read-through leaky stop codon.

ML model for EG selection

The ML model was trained on fluorescence datasets from yeast libraries (see Supplementary 1) (65, 66). The dataset included GOI-EG fusion genes for 5185 EGs (~78% of all EGs in yeast), fused with either GFP or mCherry. Fluorescence was measured across multiple time points.

Features and model architecture

The model utilized a diverse set of bioinformatic features derived from the sequences of each GOI and EG. Key feature families included the following:

1) Codon usage bias: (i) tAI: calculated for the full sequence, first 17 amino acids, and sliding windows across the GOI and EG (80); (ii) relative codon adaptation (RCA): evaluated for the GOI, EG, and sliding windows (117); (iii) codon adaptation index (CAI): assessed for the full sequence and sliding windows (81); (iv) effective number of codons: measures diversity in codon usage (118).

2) Sequence composition: (i) GC content: calculated globally and in sliding windows for both the GOI and EG (82); (ii) *k*-mer frequencies: counts of nucleotide or amino acid substrings of lengths 3 to 5; (iii) amino acid frequency correlation: Spearman correlation between GOI and EG amino acid compositions.

3) Thermodynamic properties: (i) local folding energy: calculated using ViennaRNA (119) in sliding windows (50 base pairs).

4) Chemical properties: (i) molecular weight and hydrophobicity: computed for both GOI and EG sequences (120); (ii) isoelectric point: assessed at physiological pH.

5) Translation context: (i) start codon context: features describing the nucleotide flanking regions of start codons; (ii) shifted ORF length: alternative reading frame lengths.

6) ChimeraARS score: (i) The ChimeraARS score quantifies sequence similarity to a reference set of genes with high codon usage bias. This metric enhances predictions of sequence stability beyond standard codon usage features (85).

The ML pipeline used an ensemble model combining KNN and XGB architectures (Fig. 3A). The KNN model emphasized features of the EG and provided high median performance—but was prone to failures. The XGB model emphasized features of the GOI, EG, and interaction features and provided much higher robustness to failure. Hyperparameter tuning and cross-validation were conducted using Optuna to ensure robust performance (details in the Supplementary Materials).

Feature importance in ML predictions

Feature importance was analyzed using SHAP (Shapley Additive Explanations) (121) for XGB and forward feature selection for KNN. SHAP provided a global ranking of feature contributions, while forward selection iteratively identified the most impactful features for KNN predictions. The XGB results, which were more robust, provided the primary conclusions, while KNN analysis offered additional insights where consistent or complementary.

XGB feature importance

The SHAP analysis (Fig. 3D) revealed the tAI (80) as the dominant feature, greatly outpacing all others. Variants with high tAI scores consistently showed enhanced stability and expression, underscoring its central role in optimizing translation. The influential features were as follows: (i) GC content (82) and RCA (117), both reflecting translational optimization and sequence stability; (ii) amino acid frequency correlation between the GOI and the EG, suggesting reduced metabolic burden for similar protein compositions; (iii) local folding energy (119) downstream in the GOI and near the initiation site, emphasizing the importance of avoiding unstable mRNA secondary structures; and (iv) shifted ORF length (122–124), ribosomal coverage bias score (125), and tAI in the first 17 amino acids (126–129), which further highlighted the importance of efficient translation and successful initiation.

KNN feature importance

KNN forward selection supported the importance of codon usage-related features, particularly in the EG. Key features included RCA,

tAI, and CAI (81) metrics within the EG and in specific regions, such as the first 17 amino acids. These findings complemented the more GOI-focused insights from XGB by emphasizing the role of the EG in the overall gene fusion stability.

Linkers for protein fusion

Protein misfolding remains a challenge in gene fusion strategies (130–133). While tools such as AlphaFold (134) offer an accurate prediction of protein structure, they are too computationally intensive to enable a reasonable comparison between many linkers. Rather, we used tools such as IUPred2A (84), a biophysical model, and MoreRONN (83), an ML model, to predict protein disorder profiles and assess the impact of linkers. As protein disorder profiles have a profound effect of protein folding, it is taken as a proxy variable—linkers with a smaller effect on the disorder profile are less likely to influence the protein folding and thus less likely to cause misfolding (57, 58, 135, 136). Linkers were selected to minimize disruptions to the disorder profile upon fusion, likely preserving native folding patterns of both the GOI and EG (53–59). This minimization was applied by calculating the Euclidean distance between the disorder profiles before and after fusion for the 1280 linkers in a linker database (59). The linker inducing the smallest distance was selected. Although experimental validation of the linker selection step remains pending, these predictions provide an essential foundation for mitigating misfolding risks in future applications.

Scoring and selection

Linkers were scored on the basis of the L2 distance between disorder profiles of the unfused and fused states. The optimal linker was selected to minimize disruptions to native folding patterns, preserving the functionality of both the GOI and EG. The following is the calculation

$$\Delta \mathbf{G}_\ell(x) = \mathbf{G}_\ell(x) - \mathbf{G}_{\text{unfused}}(x)$$

$$\mathcal{L}_\ell = \|\Delta \mathbf{G}_\ell(\text{EG})\|_2 + \|\Delta \mathbf{G}_\ell(\text{GOI})\|_2$$

$$\ell^* = \underset{\ell \in \mathcal{H}}{\operatorname{argmin}}(\mathcal{L}_\ell)$$

where $\mathbf{G}_\ell(x)$ is the disorder profile of gene x when fused with linker ℓ , $\mathcal{H}$ is the set of commercial linkers reported in (59), and ℓ^* is the linker yielding the minimum total disruption.

To compute disorder profiles, two complementary models were used to assign a per-residue disorder probability. IUPred2A (84) predicts disorder from estimated interresidue interaction energies, providing context-sensitive scores (0 to 1) for both short and long disordered segments. MoreRONN (83) applies a neural network trained on curated datasets to assign disorder probabilities via a sliding-window approach, enabling sensitive detection of flexible regions relevant for fusion design.

Optimizing DNA sequences for stability and expression

Sequence optimization was performed to enhance the stability and expression of the GOI-EG fusion genes using the ESO (21) pipeline. ESO integrates DNACHisel (137), a tool that optimizes genetic sequences while maintaining biological constraints. ESO has been previously validated for its ability to improve the evolutionary stability of

synthetic genes. The ESO pipeline focuses on optimizing codon usage, minimizing mRNA instability, and ensuring the long-term expression of engineered genes by accounting for evolutionary pressures.

Optimization objectives

mRNA folding optimization

It has been demonstrated that minimizing the secondary structure formation in mRNA near the translation initiation site maximizes the expression (119, 138–143). The local folding energy of the first 15 codons in the mRNA sequence was calculated using ViennaRNA (119), and codons were selected to optimize weak folding energy.

Codon usage optimization

For the rest of the sequence, codon usage was optimized to match the tRNA abundances of *S. cerevisiae*. This was achieved by calculating and optimizing the tAI (80) for the GOI, linker, and (depending on the application and relevancy) the EG, ensuring that codon usage patterns align with the host's translational machinery.

Stability enhancements

Sequences were optimized to avoid hypermutable regions (24, 110, 144–153). These optimizations ensure that synthetic genes are tailored for host-specific transcription and translation efficiencies, enhancing long-term performance in diverse applications.

Experimental validation

Fusion stability in *S. cerevisiae*

We validated the gene fusions by testing their ability to stabilize GOI expression in *S. cerevisiae*. Taking the Schuldiner lab library [N⁺ SWAp Tag (SWAT)-GFP, derived from *S. cerevisiae* BY4741 background strain] (65, 66), we selected variants from it such that the EGs are varied and representative. This was performed by performing k -means for $k = 2, \dots, 130$, with five iterations each, and selecting the 10 genes most consistently in different clusters while the closest to the cluster center. As a control, they were compared with unfused GFP, integrated into the genome by replacing *CAN1*. The *CAN1* gene encodes an arginine permease, which is not essential for yeast survival under laboratory conditions where arginine is supplemented in the medium (154–158).

Their yield was measured over 15 days of in-lab evolution, with three repeats each. The evolution experiment took place in 96-well DeepWell plates (Starlab Group, cat. no. S1896-2110). Fluorescence levels were taken as proxy for protein expression.

We first normalized this value by the optical density to control for yeast density. Furthermore, we performed this analysis for the wild-type yeast to find the inherent fluorescence of yeast. Then, we detracted this value from other measurements to isolate the fluorescence because of GFP.

Proinsulin production in yeast

To test a biotechnologically relevant application, we applied the workflow to human proinsulin. The proinsulin sequence was obtained from Novo Nordisk patented sequence for proinsulin manufacturing in yeast (patent WO2014195452A1). EGs were selected using the following criteria: (i) The EG was predicted to be in top 5% for XGB architecture. (ii) The EG was predicted to be in top 10% for KNN architecture because of less robustness. (iii) The EG length is less than 500 nt to ensure the reasonable synthesis of the combined fusion protein. (iv) The EG has no internal repeats of length >10 to ensure greater stability. (v) The EG is not ribosomal, because of many interactions of ribosomal proteins, with many points of failure. (vi) The EG does not have an induced function—it is more

sensitive to changes in expression. (vii) We chose genes from multiple cellular processes to decrease redundancy.

We constrained EG length to ≤ 500 nt to keep the total fusion construct (GOI, linker, and EG) under ~ 1 kb, optimizing DNA synthesis, cloning fidelity, and expression efficiency (159–161). This design choice applied only to synthesized constructs like proinsulin fusions; GFP constructs in Fig. 2 were assembled via SWAp-Tag recombination, where the EG length was not a limiting factor (65, 66). In our model, shorter EGs showed a weak but significant negative correlation with predicted performance (GFP Spearman $\rho = -0.055$, $P < 10^{-5}$; top-ranked EG median: 204 to 750 nt versus 1087 nt for lower ranks, $P < 0.01$). Consistently, proteome-wide analysis of 5397 verified yeast proteins revealed a small but significant negative correlation between protein abundance and length ($\log_{10}$ abundance versus length: Pearson $r = -0.15$, Spearman $\rho = -0.19$, $P < 10^{-30}$). These trends suggest that shorter EGs are particularly well suited for robust fusion design. One known cause is that shorter genes often fold efficiently and avoid the complexity associated with cotranslational folding signals. By contrast, larger, multidomain proteins are more susceptible to folding errors and aggregation, and when fused, their intricate folding pathways can interfere with the proper expression of both partners (162–165). These considerations reinforce our strategy of favoring shorter EGs in de novo synthesis.

On the basis of these criteria, *CAF20* and *ARC15* were selected. Four variants were compared, including the following: (i) the original proinsulin sequence, expressed on a high-copy-number plasmid (Prs426); (ii) a similar variant, where the proinsulin sequence was optimized by the ESO (21); (iii) a fusion gene composed of optimized proinsulin and *CAF20* expressed on a high-copy-number plasmid in a strain in which *CAF20* was deleted from the genome; and (iv) a similar variant for *ARC15*.

Each variant included an alpha leader. The linker 1fzbf was selected.

They were tested for yield over 35 days of in-lab evolution, with three repeats each. Proinsulin levels were quantified using ELISA plates according to the manufacturer's instructions (product number RAB0327) for proinsulin detection and verified by Western blotting using a proinsulin specific antibody (Sigma-Aldrich, i2018).

The constructs were analyzed through nanopore sequencing before and after the evolution experiment for a deeper understanding of the accumulated mutations. Nanopore reads were first processed using chopper (95), trimming the first 10 nt of each read and keeping reads with a length above 500. The reads were mapped with minimap2 (96) to a reference sequence, including the yeast genome and the relevant construct. Quantification of optimized variants' mRNA levels was conducted using Salmon (97), and variant calling was conducted by DeepVariant (98).

Leaky stop codon rate of read-through

Selecting and implementing a leaky stop codon are required for our design. We require a rate of read-through such that (i) the expression of the fusion protein is high enough to generate viable quantities of the fusion protein required for the cell's growth, (ii) it is the lowest value reasonable to prove more mutations deleterious, and (iii) high quantities of the GOI alone are produced, as this maximizes the output of the system.

To enable an informed selection for our stop codons, we generated different stop codons using the design principles outlined in (166). In this paper, Riboseq is conducted on all genes in *S. cerevisiae*. They created a predictive model for the rate of read-through and

found the most influential feature to be the context around the stop codon—from 3 nt before to 6 nt after, with a total of 12 nt. Further refining the model, they created a ranking of leakiness on the basis of nucleotides at each position. Using these rankings, we generated sequences predicted to be with high read-through rates.

We conducted an experiment where we implemented a fusion gene on a plasmid. This fusion gene was composed of BFP connected to the C' terminus of mCherry, with the different leaky stop codons generated. The fluorescence of BFP is proxy for the expression of both the GOI and fusion protein, and the fluorescence of mCherry is proxy for the expression of only the fusion protein.

We selected three stop codon designs such that (i) the rate of read-through, as measured in the mCherry relative fluorescence, is in the range of 0.1 to 0.25; (ii) the relative fluorescence of BFP, indicative of the expression of the isolated GOI, is greater than 0.9; and (iii) the last amino acid in proinsulin is unchanged.

We conducted another evolution experiment. Four variants were generated. In each, they followed the design of the *ARC15* variant from the previous experiment. In one, there was no stop codon, and the other three followed the stop codon designs selected.

They were tested for yield over 50 days of in-lab evolution, with three repeats each. Proinsulin levels were quantified using ELISA plates according to the manufacturer's instructions (product number RAB0327) for proinsulin detection.

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S8
Tables S1 to S4

REFERENCES AND NOTES

1. L. Zhu, Y. Zhu, Y. Zhang, Y. Li, Engineering the robustness of industrial microbes through synthetic biology. *Trends Microbiol.* **20**, 94–101 (2012).
2. M. T. Parker, A. M. Kunjapur, Deployment of engineered microbes: Contributions to the bioeconomy and considerations for biosecurity. *Health Secur.* **18**, 278–296 (2020).
3. O. S. Venturelli, R. G. Egbert, A. P. Arkin, Towards engineering biological systems in a broader context. *J. Mol. Biol.* **428**, 928–944 (2016).
4. L. Clarke, R. Kitney, Developing synthetic biology for industrial biotechnology applications. *Biochem. Soc. Trans.* **48**, 113–122 (2020).
5. W. Zhang, D. R. Nielsen, Synthetic biology applications in industrial microbiology. *Front. Microbiol.* **5**, fmicb.2014.00451 (2014).
6. L. Katz, Y. Y. Chen, R. Gonzalez, T. C. Peterson, H. Zhao, R. H. Baltz, Synthetic biology advances and applications in the biotechnology industry: A perspective. *J. Ind. Microbiol. Biotechnol.* **45**, 449–461 (2018).
7. B. W. Biggs, H. S. Alper, B. F. Pfeiffer, K. E. J. Tyo, C. N. S. Santos, P. K. Ajikumar, G. Stephanopoulos, Enabling commercial success of industrial biotechnology. *Science* **374**, 1563–1565 (2021).
8. K. L. Arnolds, L. R. Dahlin, L. Ding, C. Wu, J. Yu, W. Xiong, C. Zuniga, Y. Suzuki, K. Zengler, J. G. Linger, M. T. Guarnieri, Biotechnology for secure biocontainment designs in an emerging bioeconomy. *Curr. Opin. Biotechnol.* **71**, 25–31 (2021).
9. C. A. Voigt, Synthetic biology 2020–2030: Six commercially-available products that are changing our world. *Nat. Commun.* **11**, 6379 (2020).
10. E. T. Wurtzel, C. E. Vickers, A. D. Hanson, A. H. Millar, M. Cooper, K. P. Voss-Fels, P. I. Nikel, T. J. Erb, Revolutionizing agriculture with synthetic biology. *Nat. Plants* **5**, 1207–1210 (2019).
11. J. C. Mortimer, Plant synthetic biology could drive a revolution in biofuels and medicine. *Exp. Biol. Med.* **244**, 323–331 (2019).
12. J. Claesen, M. A. Fischbach, Synthetic microbes as drug delivery systems. *ACS Synth. Biol.* **4**, 358–364 (2015).
13. S. Slomovic, K. Pardee, J. J. Collins, Synthetic biology devices for in vitro and in vivo diagnostics. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 14429–14435 (2015).
14. X. Tan, J. H. Letendre, J. J. Collins, W. W. Wong, Synthetic biology in the clinic: Engineering vaccines, diagnostics, and therapeutics. *Cell* **184**, 881–898 (2021).

15. J. L. Chlebek, S. P. Leonard, C. Kang-Yun, M. C. Yung, D. P. Ricci, Y. Jiao, D. M. Park, Prolonging genetic circuit stability through adaptive evolution of overlapping genes. *Nucleic Acids Res.* **51**, 7094–7108 (2023).
16. S. C. Sleight, H. M. Sauro, Visualization of evolutionary stability dynamics and competitive fitness of *Escherichia coli* engineered with randomized multigene circuits. *ACS Synth. Biol.* **2**, 519–528 (2013).
17. S. C. Sleight, B. A. Bartley, J. A. Lieviant, H. M. Sauro, Designing and engineering evolutionary robust genetic circuits. *J. Biol. Eng.* **4**, 12 (2010).
18. S. L. Nuismer, N. C. Layman, A. J. Redwood, B. Chan, J. J. Bull, Methods for measuring the evolutionary stability of engineered genomes to improve their longevity. *Synth. Biol.* **6**, ysab018 (2021).
19. M. Arbel-Groissman, I. Menuhin-Gruman, D. Naki, S. Bergman, T. Tuller, Fighting the battle against evolution: Designing genetically modified organisms for evolutionary stability. *Trends Biotechnol.* **41**, 1518–1531 (2023).
20. M. Arbel-Groissman, I. Menuhin-Gruman, H. Yehezkeili, D. Naki, S. Bergman, Y. Udi, T. Tuller, The causes for genomic instability and how to try and reduce them through rational design of synthetic DNA. *Methods Mol. Biol.* **2760**, 371–392 (2024).
21. I. Menuhin-Gruman, M. Arbel, N. Amitay, K. Sionov, D. Naki, I. Katzir, O. Edgar, S. Bergman, T. Tuller, Evolutionary Stability Optimizer (ESO): A novel approach to identify and avoid mutational hotspots in DNA sequences while maintaining high expression levels. *ACS Synth. Biol.* **11**, 1142–1151 (2022).
22. S. Yang, S. C. Sleight, H. M. Sauro, Rationally designed bidirectional promoter improves the evolutionary stability of synthetic genetic circuits. *Nucleic Acids Res.* **41**, e33 (2013).
23. D. Ingram, G. B. Stan, Modelling genetic stability in engineered cell populations. *Nat. Commun.* **14**, 3471 (2023).
24. B. R. Jack, S. P. Leonard, D. M. Mishler, B. A. Renda, D. Leon, G. A. Suárez, J. E. Barrick, Predicting the genetic stability of engineered DNA sequences with the EFM calculator. *ACS Synth. Biol.* **4**, 939–943 (2014).
25. R. L. Williams, R. M. Murray, Integrase-mediated differentiation circuits improve evolutionary stability of burdensome and toxic functions in *E. coli*. *Nat. Commun.* **13**, 6822 (2022).
26. H. I. Son, A. Weiss, L. You, Design patterns for engineering genetic stability. *Curr. Opin. Biomed. Eng.* **19**, 100297 (2021).
27. I. Frumkin, D. Schirman, A. Rotman, F. Li, L. Zahavi, E. Mordret, O. Asraf, S. Wu, S. F. Levy, Y. Pilpel, Gene architectures that minimize cost of gene expression. *Mol. Cell* **65**, 142–153 (2017).
28. B. A. Renda, M. J. Hammerling, J. E. Barrick, Engineering reduced evolutionary potential for synthetic biology. *Mol. Biosyst.* **10**, 1668–1678 (2014).
29. A. Willemssen, M. P. Zwart, On the stability of sequences inserted into viral genomes. *Virus Evol.* **5**, vez045 (2019).
30. P. Rugbjerg, N. Myling-Petersen, A. Porse, K. Sarup-Lytzen, M. O. A. Sommer, Diverse genetic error modes constrain large-scale bio-based production. *Nat. Commun.* **9**, 787 (2018).
31. B. R. Glick, Metabolic load and heterologous gene expression. *Biotechnol. Adv.* **13**, 247–261 (1995).
32. O. Borkowski, C. Bricio, M. Murgiano, B. Rothschild-Mancinelli, G. B. Stan, T. Ellis, Cell-free prediction of protein expression costs for growing cells. *Nat. Commun.* **9**, 1457 (2018).
33. J. F. Görgens, W. H. Van Zyl, J. H. Knoetze, B. Hahn-Hägerdal, The metabolic burden of the PGK1 and ADH2 promoter systems for heterologous xylanase production by *saccharomyces cerevisiae* in defined medium. *Biotechnol. Bioeng.* **73**, 238–245 (2001).
34. Q. Liu, J. Schumacher, X. Wan, C. Lou, B. Wang, Orthogonality and burdens of heterologous and gate gene circuits in *E. coli*. *ACS Synth. Biol.* **7**, 553–564 (2018).
35. U. Sauer, Evolutionary engineering of industrially important microbial phenotypes. *Adv. Biochem. Eng. Biotechnol.* **73**, 129–169 (2001).
36. T. Frei, F. Cella, F. Tedeschi, J. Gutiérrez, G. B. Stan, M. Khammash, V. Siciliano, Characterization and mitigation of gene expression burden in mammalian cells. *Nat. Commun.* **11**, 4641 (2020).
37. S. M. Brooks, H. S. Alper, Applications, challenges, and needs for employing synthetic biology beyond the lab. *Nat. Commun.* **12**, 1390 (2021).
38. G. N. Mandel, G. E. Marchant, The living regulatory challenges of synthetic biology. *Iowa L. Rev.* **100**, 155–200 (2014).
39. B. Rager-Zisman, Ethical and regulatory challenges posed by synthetic biology. *Perspect. Biol. Med.* **55**, 590–607 (2012).
40. A. Hossain, E. Lopez, S. M. Halper, D. P. Cetnar, A. C. Reis, D. Strickland, E. Klavins, H. M. Salis, Automated design of thousands of nonrepetitive parts for engineering stable genetic systems. *Nat. Biotechnol.* **38**, 1466–1475 (2020).
41. P. Cárdenas, Designing for durability: New tools to build stable, non-repetitive DNA. *Synth. Biol.* **5**, ysaa016 (2020).
42. M. J. Liao, M. O. Din, L. Tsimring, J. Hasty, Rock-paper-scissors: Engineered population dynamics increase genetic stability. *Science* **365**, 1045–1049 (2019).
43. N. Dalchau, M. J. Smith, S. Martin, J. R. Brown, S. Emmott, A. Phillips, Towards the rational design of synthetic cells with prescribed population dynamics. *J. R. Soc. Interface* **9**, 2883–2898 (2012).
44. H. Song, S. Payne, C. Tan, L. You, Programming microbial population dynamics by engineered cell-cell communication. *Biotechnol. J.* **6**, 837–849 (2011).
45. A. L. Decrulle, A. Frénoy, T. A. Meiller-Legrand, A. Bernheim, C. Lotton, A. Gutierrez, A. B. Lindner, Engineering gene overlaps to sustain genetic constructs in vivo. *PLoS Comput. Biol.* **17**, e1009475 (2021).
46. T. Blazejewski, H. I. Ho, H. H. Wang, Synthetic sequence entanglement augments stability and containment of genetic information in cells. *Science* **365**, 595–598 (2019).
47. S. Raman, J. K. Rogers, N. D. Taylor, G. M. Church, Evolution-guided optimization of biosynthetic pathways. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 17803–17808 (2014).
48. P. Rugbjerg, K. Sarup-Lytzen, M. Nagy, M. O. A. Sommer, Synthetic addiction extends the productive life time of engineered *Escherichia coli* populations. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2347–2352 (2018).
49. A. C. Stiebler, J. Freitag, K. O. Schink, T. Stehlik, B. A. M. Tillmann, J. Ast, M. Bölker, Ribosomal readthrough at a short UGA stop codon context triggers dual localization of metabolic enzymes in fungi and animals. *PLoS Genet.* **10**, e1004685 (2014).
50. J. M. L. Ho, C. A. Miller, S. E. Parks, J. R. Mattia, M. R. Bennett, A suppressor tRNA-mediated feedforward loop eliminates leaky gene expression in bacteria. *Nucleic Acids Res.* **49**, e25 (2021).
51. O. D. Caspari, Introduction of a leaky stop codon as molecular tool in *Chlamydomonas reinhardtii*. *PLOS ONE* **15**, e0237405 (2020).
52. K. M. Keeling, J. Lanier, M. Du, J. Salas-Marco, L. Gao, A. Kaenjak-Angeletti, D. M. Bedwell, Leaky termination at premature stop codons antagonizes nonsense-mediated mRNA decay in *S. cerevisiae*. *RNA* **10**, 691–703 (2004).
53. A. V. Uversky, V. N. Uversky, "Amino acid code for protein folding, misfolding, and non-folding" in *Amino Acids, Peptides and Proteins* (The Royal Society of Chemistry, 2014), vol. 39, pp. 192–236.
54. J. Y. Yang, M. Yang, Predicting protein disorder by analyzing amino acid sequence. *BMC Genomics* **9**, S8 (2008).
55. N. Rajasekaran, S. Gopi, A. Narayan, A. N. Naganathan, Quantifying protein disorder through measures of excess conformational entropy. *J. Phys. Chem. B* **120**, 4341–4350 (2016).
56. J. A. Raskatov, D. B. Teplow, Using chirality to probe the conformational dynamics and assembly of intrinsically disordered amyloid proteins. *Sci. Rep.* **7**, 12433 (2017).
57. V. Receveur-Bréhot, J. M. Bourhis, V. N. Uversky, B. Canard, S. Longhi, Assessing protein disorder and induced folding. *Proteins* **62**, 24–45 (2006).
58. A. Toto, F. Malagrino, L. Visconti, F. Troilo, L. Pagano, M. Brunori, P. Jemth, S. Gianni, Templated folding of intrinsically disordered proteins. *J. Biol. Chem.* **295**, 6586–6593 (2020).
59. R. A. George, J. Heringa, An analysis of protein domain linkers: Their classification and role in protein folding. *Protein Eng.* **15**, 871–879 (2002).
60. J. S. Klein, S. Jiang, R. P. Galimidi, J. R. Keffee, P. J. Bjorkman, L. Regan, Design and characterization of structured protein linkers with differing flexibilities. *Protein Eng. Des. Sel.* **27**, 325–330 (2014).
61. V. Presnyak, N. Alhusaini, Y. H. Chen, S. Martin, N. Morris, N. Kline, S. Olson, D. Weinberg, K. E. Baker, B. R. Graveley, J. Collier, Codon optimality is a major determinant of mRNA stability. *Cell* **160**, 1111–1124 (2015).
62. S. G. Medina-Muñoz, G. Kushawah, L. A. Castellano, M. Diez, M. L. DeVore, M. J. B. Salazar, A. O. Goldmani, Crosstalk between codon optimality and cis-regulatory elements dictates mRNA stability. *Genome Biol.* **22**, 14 (2021).
63. H. Zhang, L. Zhang, A. Lin, C. Xu, Z. Li, K. Liu, B. Liu, X. Ma, F. Zhao, H. Jiang, C. Chen, H. Shen, H. Li, D. H. Mathews, Y. Zhang, L. Huang, Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature* **621**, 396–403 (2023).
64. I. Jonassen, I. G. Clausen, E. B. Jensen, A. Svendsen, US5324641 (1994); <https://patents.google.com/patent/US5324641A/en>.
65. I. Yofe, U. Weill, M. Meurer, S. Chuartzman, E. Zalckvar, O. Goldman, S. Ben-Dor, C. Schütze, N. Wiedemann, M. Knop, A. Khmelinskii, M. Schuldiner, One library to make them all: Streamlining the creation of yeast libraries via a SWAp-Tag strategy. *Nat. Methods* **13**, 371–378 (2016).
66. U. Weill, I. Yofe, E. Sass, B. Stynen, D. Davidi, J. Natarajan, R. Ben-Menachem, Z. Avihou, O. Goldman, N. Harpaz, S. Chuartzman, K. Kniazhev, K. Knoblich, J. Laborenz, F. Boos, J. Kowarzyk, S. Ben-Dor, E. Zalckvar, J. M. Herrmann, R. A. Rachubinski, O. Pines, D. Rapaport, S. W. Michnick, E. D. Levy, M. Schuldiner, Genome-wide SWAp-Tag yeast libraries for proteome exploration. *Nat. Methods* **15**, 617–622 (2018).
67. M. Chalfie, Green fluorescent protein as a marker for gene expression. *Trends Genet.* **10**, 802–805 (1994).
68. G. S. Waldo, B. M. Standish, J. Berendzen, T. C. Terwilliger, Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol.* **17**, 691–695 (1999).
69. M. G. Romei, S. G. Boxer, Split green fluorescent proteins: Scope, limitations, and outlook. *Annu. Rev. Biophys.* **48**, 19–44 (2019).
70. J. Kong, Y. Wang, W. Qi, M. Huang, R. Su, Z. He, Green fluorescent protein inspired fluorophores. *Adv. Colloid. Interface Sci.* **285**, 102286 (2020).
71. S. R. Kain, Green fluorescent protein (GFP): Applications in cell-based assays for drug discovery. *Drug Disc. Today* **4**, 304–312 (1999).

72. M. Meurer, Y. Duan, E. Sass, I. Kats, K. Herbst, B. C. Buchmuller, V. Dederer, F. Huber, D. Kirrmaier, M. Štefl, K. Van Laer, T. P. Dick, M. K. Lemberg, A. Khmelinskii, E. D. Levy, M. Knop, Genome-wide C-SWAT library for high-throughput yeast genome tagging. *Nat. Methods* **15**, 598–600 (2018).
73. A. R. Hieb, S. D'Arcy, M. A. Kramer, A. E. White, K. Luger, Fluorescence strategies for high-throughput quantification of protein interactions. *Nucleic Acids Res.* **40**, e33 (2012).
74. J. A. J. Arpino, P. J. Rizkallah, D. D. Jones, Structural and dynamic changes associated with beneficial engineered single-amino-acid deletion mutations in enhanced green fluorescent protein. *Acta Crystallogr. D Biol. Crystallogr.* **70**, 2152–2162 (2014).
75. L. Pillai-Kastoori, S. Heaton, S. D. Shiflett, A. C. Roberts, A. Solache, A. R. Schutz-Geschwender, Antibody validation for Western blot: By the user, for the user. *J. Biol. Chem.* **295**, 926–939 (2020).
76. J. E. Gilda, R. Ghosh, J. X. Cheah, T. M. West, S. C. Bodine, A. V. Gomes, Western blotting inaccuracies with unverified antibodies: Need for a Western Blotting Minimal Reporting Standard (WBMRS). *PLOS ONE* **10**, e0135392 (2015).
77. S. C. Taylor, L. K. Rosselli-Murai, B. Crobbedu, I. Plante, A critical path to producing high quality, reproducible data from quantitative western blot experiments. *Sci. Rep.* **12**, 17599 (2022).
78. S. C. Taylor, A. Posch, The design of a quantitative western blot experiment. *BioMed. Res. Int.* **2014**, 361590 (2014).
79. H. P. J. Buermans, J. T. den Dunnen, Next generation sequencing technology: Advances and applications. *BBA Mol. Basis Dis.* **1842**, 1932–1941 (2014).
80. M. dos Reis, R. Savva, L. Wernisch, Solving the riddle of codon usage preferences: A test for translational selection. *Nucleic Acids Res.* **32**, 5036–5044 (2004).
81. P. M. Sharp, W. H. Li, The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
82. D. A. Kiktev, Z. Sheng, K. S. Lobachev, T. D. Petes, GC content elevates mutation and recombination rates in the yeast *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E7109–E7118 (2018).
83. Z. R. Yang, R. Thomson, P. McNeil, R. M. Esnouf, RONN: The bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* **21**, 3369–3376 (2005).
84. B. Mészáros, G. Erdős, Z. Dosztányi, IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337 (2018).
85. H. Zur, T. Tuller, Exploiting hidden information interleaved in the redundancy of the genetic code without prior knowledge. *Bioinformatics* **31**, 1161–1168 (2015).
86. E. Fix, J. L. Hodges, Discriminatory analysis. Nonparametric discrimination: Consistency properties. *Int. Stat. Rev.* **57**, 238–247 (1989).
87. T. Chen, C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2016), pp. 785–794.
88. C. H. D. Moor, M. Jansen, E. J. Bonte, A. A. M. Thomas, J. S. Sussenbach, J. L. Van Den Brande, Influence of the four leader sequences of the human insulin-like-growth-factor-2 mRNAs on the expression of reporter genes. *Eur. J. Biochem.* **226**, 1039–1047 (1994).
89. T. Kjeldsen, A. F. Pettersson, M. Hach, I. Diers, S. Havelund, P. H. Hansen, A. S. Andersen, Synthetic leaders with potential BiP binding mediate high-yield secretion of correctly folded insulin precursors from *Saccharomyces cerevisiae*. *Protein Expr. Purif.* **9**, 331–336 (1997).
90. T. Kjeldsen, Yeast secretory expression of insulin precursors. *Appl. Microbiol. Biotechnol.* **54**, 277–286 (2000).
91. C. K. Min, Y. J. Son, C. K. Kim, S. J. Park, J. W. Lee, Increased expression, folding and enzyme reaction rate of recombinant human insulin by selecting appropriate leader peptide. *J. Biotechnol.* **151**, 350–356 (2011).
92. L. Thim, M. T. Hansen, K. Norris, I. Hoegh, E. Boel, J. Forstrom, G. Ammerer, N. P. Fiil, Secretion and processing of insulin precursors in yeast. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 6766–6770 (1986).
93. H. Teerink, H. O. Voorma, A. A. M. Thomas, The human insulin-like growth factor II leader 1 contains an internal ribosomal entry site. *Biochim. Biophys. Acta* **1264**, 403–408 (1995).
94. H. Yang, M. L. Adamo, A. P. Koval, M. C. McGuinness, H. Ben-Hur, Y. Yang, D. LeRoith, C. T. Roberts, Alternative leader sequences in insulin-like growth factor I mRNAs modulate translational efficiency and encode multiple signal peptides. *Mol. Endocrinol.* **9**, 1380–1395 (1995).
95. W. De Coster, R. Rademakers, NanoPack2: Population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, btad311 (2023).
96. H. Li, New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
97. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
98. R. Poplin, P. C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, S. S. Gross, L. Dorfman, C. Y. McLean, M. A. Deprieto, A universal snp and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
99. D. A. Drummond, C. O. Wilke, Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352 (2008).
100. A. A. Komar, E. Samatova, M. V. Rodnina, Translation rates and protein folding. *J. Mol. Biol.* **436**, 168384 (2024).
101. A. E. Hirsh, H. B. Fraser, D. P. Wall, Adjusting for selection on synonymous sites in estimates of evolutionary distance. *Mol. Biol. Evol.* **22**, 174–177 (2005).
102. W. Lv, J. Zheng, M. Luan, M. Shi, H. Zhu, M. Zhang, H. Lv, Z. Shang, L. Duan, R. Zhang, Y. Jiang, Comparing the evolutionary conservation between human essential genes, human orthologs of mouse essential genes and human housekeeping genes. *Brief. Bioinform.* **16**, 922–931 (2015).
103. M. M. Albà, J. Castresana, Inverse relationship between evolutionary rate and age of mammalian genes. *Mol. Biol. Evol.* **22**, 598–606 (2005).
104. H. Zhang, Z. Lyu, Y. Fan, C. R. Evans, K. W. Barber, K. Banerjee, O. A. Igoshin, J. Rinehart, J. Ling, Metabolic stress promotes stop-codon readthrough and phenotypic heterogeneity. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 22167–22172 (2020).
105. Y. Fan, C. R. Evans, K. W. Barber, K. Banerjee, K. J. Weiss, W. Margolin, O. A. Igoshin, J. Rinehart, J. Ling, Heterogeneity of stop codon readthrough in single bacterial cells and implications for population fitness. *Mol. Cell* **67**, 826–836.e5 (2017).
106. I. Jungreis, M. F. Lin, R. Spokony, C. S. Chan, N. Negre, A. Victorson, K. P. White, M. Kellis, Evidence of abundant stop codon readthrough in *Drosophila* and other metazoa. *Genome Res.* **21**, 2069–2113 (2011).
107. G. Loughran, M. Y. Chou, I. P. Ivanov, I. Jungreis, M. Kellis, A. M. Kiran, P. V. Baranov, J. F. Atkins, Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Res.* **42**, 8928–8938 (2014).
108. L. E. Manjunath, A. Singh, S. Som, S. M. Eswarappa, Mammalian proteome expansion by stop codon readthrough. *Wiley Interdiscip. Rev. RNA* **14**, e1739 (2023).
109. M. Wang, K. Zhang, V. Ngo, C. Liu, S. Fan, J. W. Whitaker, Y. Chen, R. Ai, Z. Chen, J. Wang, L. Zheng, W. Wang, Identification of DNA motifs that regulate DNA methylation. *Nucleic Acids Res.* **47**, 6753–6768 (2019).
110. W. H. Lin, E. Kussell, Evolutionary pressures on simple sequence repeats in prokaryotic coding regions. *Nucleic Acids Res.* **40**, 2399–2413 (2012).
111. E. Li, Y. Zhang, DNA methylation in mammals. *Cold Spring Harb. Perspect. Biol.* **6**, 676–707 (2014).
112. K. M. Glastad, B. G. Hunt, S. V. Yi, M. A. D. Goodisman, DNA methylation in insects: On the brink of the epigenomic era. *Insect Mol. Biol.* **20**, 553–565 (2011).
113. M. Curradi, A. Izzo, G. Badaracco, N. Landsberger, Molecular mechanisms of gene silencing mediated by DNA methylation. *Mol. Cell. Biol.* **22**, 3157–3173 (2002).
114. J. Newell-Price, A. J. L. Clark, P. King, DNA methylation and silencing of gene expression. *Trends Endocrinol. Metab.* **11**, 142–148 (2000).
115. S. Rajeevkumar, P. Anunanthini, R. Sathishkumar, Epigenetic silencing in transgenic plants. *Front. Plant Sci.* **6**, 693 (2015).
116. S. I. S. Grewal, D. Moazed, Heterochromatin and epigenetic control of gene expression. *Science* **301**, 798–802 (2003).
117. J. M. Fox, I. Erill, Relative codon adaptation: A generic codon bias index for prediction of gene expression. *DNA Res.* **17**, 185–196 (2010).
118. F. Wright, The “effective number of codons” used in a gene. *Gene* **87**, 23–29 (1990).
119. R. Lorenz, S. H. Bernhart, C. H. zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker, ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
120. J. R. Lobry, C. Gautier, Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res.* **22**, 3174–3180 (1994).
121. S. M. Lundberg, S. I. Lee, “A unified approach to interpreting model predictions,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Curran Associates Inc., 2017), pp. 4768–4777.
122. J. P. Ferreira, K. W. Overton, C. L. Wang, Tuning gene expression with synthetic upstream open reading frames. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 11284–11289 (2013).
123. J. G. Canestrari, E. Lasek-Nesselquist, A. Upadhyay, M. Rofaeil, M. M. Champion, J. T. Wade, K. M. Derbyshire, T. A. Gray, Polycysteine-encoding leaderless short ORFs function as cysteine-responsive attenuators of operonic gene expression in mycobacteria. *Mol. Microbiol.* **114**, 93–108 (2020).
124. A. Moustakas, T. S. Sonstegard, P. B. Hackett, Alterations of the three short open reading frames in the Rous sarcoma virus leader RNA modulate viral replication and gene expression. *J. Virol.* **67**, 4337–4349 (1993).
125. U. Roymondal, S. Das, S. Sahoo, Predicting gene expression level from relative codon usage bias: An application to *Escherichia coli* genome. *DNA Res.* **16**, 13–30 (2009).
126. S. Grünert, R. J. Jackson, The immediate downstream codon strongly influences the efficiency of utilization of eukaryotic translation initiation codons. *EMBO J.* **13**, 3618–3630 (1994).
127. H. Miyasaka, The positive relationship between codon usage bias and translation initiation AUG context in *Saccharomyces cerevisiae*. *Yeast* **15**, 633–637 (1999).

128. T. Sato, M. Terabe, H. Watanabe, T. Gojobori, C. Hori-Takemoto, K. I. Miura, Codon and base biases after the initiation codon of the open reading frames in the *Escherichia coli* genome and their influence on the translation efficiency. *J. Biochem.* **129**, 851–860 (2001).
129. T. Tuller, A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, Y. Pilpel, An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344–354 (2010).
130. Y. Rong, S. I. Jensen, K. Lindorff-Larsen, A. T. Nielsen, Folding of heterologous proteins in bacterial cell factories: Cellular mechanisms and engineering strategies. *Biotechnol. Adv.* **63**, 108079 (2023).
131. C. Kurland, J. Gallant, Errors of heterologous protein expression. *Curr. Opin. Biotechnol.* **7**, 489–493 (1996).
132. A. Ohno, J. I. Maruyama, T. Nemoto, M. Arioka, K. Kitamoto, A carrier fusion significantly induces unfolded protein response in heterologous protein production by *Aspergillus oryzae*. *Appl. Microbiol. Biotechnol.* **92**, 1197–1206 (2011).
133. E. V. Grosfeld, A. Y. Beizer, A. A. Dergalev, M. O. Agaphonov, A. I. Alexandrov, Fusion of Hsp70 to GFP impairs its function and causes formation of misfolded protein deposits under mild stress in yeast. *Int. J. Mol. Sci.* **24**, 12758 (2023).
134. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
135. K. Sugase, H. J. Dyson, P. E. Wright, Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* **447**, 1021–1025 (2007).
136. B. Mészáros, L. Dobson, E. Fichó, G. E. Tusnády, Z. Dosztányi, I. Simon, Sequential, structural and functional properties of protein complexes are defined by how folding and binding intertwine. *J. Mol. Biol.* **431**, 4408–4428 (2019).
137. V. Zulkower, S. Rosser, DNA Chisel, a versatile sequence optimizer. *Bioinformatics* **36**, 4508–4509 (2020).
138. T. Tuller, Y. Y. Waldman, M. Kupiec, E. Ruppin, Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 3645–3650 (2010).
139. J. Yin, L. Bao, H. Tian, X. Gao, W. Yao, Quantitative relationship between the mRNA secondary structure of translational initiation region and the expression level of heterologous protein in *Escherichia coli*. *J. Ind. Microbiol. Biotechnol.* **43**, 97–102 (2016).
140. M. H. De Smit, J. Van Duin, Control of translation by mRNA secondary structure in *Escherichia coli*. A quantitative analysis of literature data. *J. Mol. Biol.* **244**, 144–150 (1994).
141. A. Espah Borujeni, H. M. Salis, Translation Initiation is controlled by RNA folding kinetics via a ribosome drafting mechanism. *J. Am. Chem. Soc.* **138**, 7016–7023 (2016).
142. M. Peeri, T. Tuller, High-resolution modeling of the selection on local mRNA folding strength in coding sequences across the tree of life. *Genome Biol.* **21**, 63 (2020).
143. W. Gu, T. Zhou, C. O. Wilke, A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput. Biol.* **6**, e1000664 (2010).
144. H. Karaoglu, C. M. Y. Lee, W. Meyer, Survey of simple sequence repeats in completed fungal genomes. *Mol. Biol. Evol.* **22**, 639–649 (2005).
145. D. Tautz, C. Schlötterer, Simple sequences. *Curr. Opin. Genet. Dev.* **4**, 832–837 (1994).
146. D. Tautz, Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res.* **17**, 6463–6471 (1989).
147. P. H. Oliveira, F. Lemos, G. A. Monteiro, D. M. F. Prazeres, Recombination frequency in plasmid DNA containing direct repeats-predictive correlation with repeat and intervening sequence length. *Plasmid* **60**, 159–165 (2008).
148. N. Phadnis, R. A. Sia, E. A. Sia, Analysis of repeat-mediated deletions in the mitochondrial genome of *Saccharomyces cerevisiae*. *Genetics* **171**, 1549–1559 (2005).
149. C. J. Saveson, S. T. Lovett, Enhanced deletion formation by aberrant DNA replication in *Escherichia coli*. *Genetics* **146**, 457–470 (1997).
150. R. E. Lenski, T. T. Nguyen, Stability of recombinant DNA and its effects on fitness. *Trends Ecol. Evol.* **3**, S18–S20 (1988).
151. X. Li, W. D. Heyer, Homologous recombination in DNA repair and DNA damage tolerance. *Cell Res.* **18**, 99–113 (2008).
152. S. T. Lovett, Encoded errors: Mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol. Microbiol.* **52**, 1243–1253 (2004).
153. I. B. Rogozin, Y. I. Pavlov, Theoretical analysis of mutation hotspots and their DNA sequence context specificity. *Mutat. Res. Rev. Mutat. Res.* **544**, 65–85 (2003).
154. A. Nishimura, K. Nakagami, K. Kan, F. Morita, H. Takagi, Arginine inhibits *Saccharomyces cerevisiae* biofilm formation by inducing endocytosis of the arginine transporter Can1. *Biosci. Biotechnol. Biochem.* **86**, 1300–1307 (2022).
155. P. Zhang, X. Hu, Metabolic engineering of arginine permeases to reduce the formation of urea in *Saccharomyces cerevisiae*. *World J. Microbiol. Biotechnol.* **34**, 47 (2018).
156. M. Ahmad, H. Bussey, Yeast arginine permease: Nucleotide sequence of the CAN1 gene. *Curr. Genet.* **10**, 587–592 (1986).
157. B. Regenber, M. C. Kielland-Brandt, Amino acid residues important for substrate specificity of the amino acid permeases Can1p and Gnp1p in *Saccharomyces cerevisiae*. *Yeast* **18**, 1429–1440 (2001).
158. P. A. Fantes, J. Creanor, Canavanine resistance and the mechanism of arginine uptake in the fission yeast *Schizosaccharomyces pombe*. *J. Gen. Microbiol.* **130**, 3265–3273 (1984).
159. J. Fu, X. Bian, S. Hu, H. Wang, F. Huang, P. M. Seibert, A. Plaza, L. Xia, R. Müller, A. F. Stewart, Y. Zhang, Full-length RecE enhances linear-linear homologous recombination and facilitates direct cloning for bioprospecting. *Nat. Biotechnol.* **30**, 440–446 (2012).
160. H. Okayama, P. Berg, High-efficiency cloning of full-length cDNA. *Mol. Cell. Biol.* **2**, 161–170 (1982).
161. K. Volkenborn, L. Kuschmierz, N. Benz, P. Lenz, A. Knapp, K. E. Jaeger, The length of ribosomal binding site spacer sequence controls the production yield for intracellular and secreted proteins by *Bacillus subtilis*. *Microb. Cell Fact.* **19**, 154 (2020).
162. N. Rajasekaran, C. M. Kaiser, Co-translational folding of multi-domain proteins. *Front. Mol. Biosci.* **9**, 869027 (2022).
163. A. Beygmoradi, A. Homaei, R. Hemmati, P. Fernandes, Recombinant protein expression: Challenges in production and folding related matters. *Int. J. Biol. Macromol.* **233**, 123407 (2023).
164. J. Marino, G. Von Heijne, R. Beckmann, Small protein domains fold inside the ribosome exit tunnel. *FEBS Lett.* **590**, 655–660 (2016).
165. S. W. Englander, L. Mayne, The case for defined protein folding pathways. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 8253–8258 (2017).
166. K. Mangkalaphiban, F. He, R. Ganesan, C. Wu, R. Baker, A. Jacobson, Transcriptome-wide investigation of stop codon readthrough in *Saccharomyces cerevisiae*. *PLOS Genet.* **17**, e1009538 (2021).

Acknowledgments: We thank M. Schuldiner's lab for providing us with SWAp-Tag strains and data. We thank D. Dovrat for helpful discussions and comments. **Funding:** S.B. is supported by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. The study was also supported by a donation from Lonza Bioscience to the TAU IGEN team and by a grant from the Israeli innovation authority. **Author contributions:** I.M.-G.: conceptualization, methodology, resources, data curation, software, formal analysis, visualization, investigation, validation, writing—original draft, and writing—review and editing. T.T.: conceptualization, methodology, resources, funding acquisition, software, formal analysis, visualization, investigation, validation, supervision, project administration, writing—original draft, and writing—review and editing. M.A.-G.: conceptualization, methodology, resources, investigation, validation, and writing—original draft. S.B.: data curation, software, formal analysis, visualization, writing—original draft, and writing—review and editing. D.N.: data curation, resources, investigation, validation, and writing—original draft. **Competing interests:** Ramot (TAU TTO) submitted two patent applications related to this paper. The authors of the patents are identical to the authors of this paper. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. All code, processed data, and materials necessary to evaluate the conclusions of this study are available in a permanent Zenodo repository at <https://doi.org/10.5281/zenodo.16959725>. The codebase and accompanying README are also hosted for convenience at https://github.com/itamar-menuhin/STABLES_code.

Submitted 3 March 2025
 Accepted 29 August 2025
 Published 1 October 2025
 10.1126/sciadv.adx0796

AI-directed gene fusing prolongs the evolutionary half-life of synthetic gene circuits

Itamar Menuhin-Gruman, Matan Arbel-Groissman, Doron Naki, Shaked Bergman, and Tamir Tuller

Sci. Adv. **11** (40), eadx0796. DOI: 10.1126/sciadv.adx0796

View the article online

<https://www.science.org/doi/10.1126/sciadv.adx0796>

Permissions

<https://www.science.org/help/reprints-and-permissions>

Use of this article is subject to the [Terms of service](#)

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.

Copyright © 2025 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).