Roots *et al. Journal of Biological Engineering* (2025) 19:50
https://doi.org/10.1186/s13036-025-00521-z

Journal of Biological Engineering

RESEARCH Open Access

Codon usage modulates the relationship between the burden and yield of protein overexpression

Cameron T. Roots[1†], Alexis M. Hill[2†], Claus O. Wilke[2] and Jeffrey E. Barrick[1*]

Abstract

Background Excess utilization of translational resources is a critical source of burden on cells engineered to overexpress exogenous proteins. To improve translational efficiency, researchers often modify codon usage in an exogenous gene to more closely match the composition of a host organism's highly expressed genes. Despite empirical data showing the benefits of codon optimization, little is known about the quantitative relationships between codon usage, protein yield, and the burden imposed on a host cell by protein overexpression.

Results We develop and experimentally evaluate a stochastic gene expression model that considers the impact of codon usage bias on the availability of ribosomes and different tRNAs in a cell. In agreement with other studies, our model shows that increasing exogenous protein expression decreases production of native cellular proteins in a linear fashion. We also find that the slope of this relationship is modulated by how well the codon usage bias of the exogenous gene and the host's genes match. Lastly, our model predicts that an overoptimization domain exists where further increasing usage of optimal codons worsens yield and burden. We test our model by expressing sfGFP and mCherry2 from constructs that have a wide range of codon optimization levels in *Escherichia coli*. The results agree with our model, including for an mCherry2 gene sequence that appears to less efficiently express this gene due to codon overoptimization.

Conclusions Our model reproduces experimentally observed relationships between codon usage bias, gene expression, and burden for overexpressed proteins. Furthermore, it suggests that more nuanced recoding strategies that seek to match a host's overall codon usage bias are less burdensome and will lead to greater protein yields compared to strategies that simply maximize usage of optimal codons. Increasing the level of mechanistic detail in gene expression models can lead to insights that allow researchers to engineer more optimal cellular systems.

Introduction

Bacterial cells have evolved metabolic pathways and gene regulatory networks that achieve high fitness in their natural environments. When bacteria are genetically engineered, their new functions can impact the availability of cellular resources, induce stress/toxicity, or otherwise interfere with growth and survival [1–4]. The fitness cost of adding an engineered DNA construct to a host cell is referred to as its burden. Constructs with high burden can be evolutionary unstable or even unclonable if cells with mutations that alleviate that burden rapidly arise

†Cameron T. Roots and Alexis M. Hill contributed equally to this work.

*Correspondence:
Jeffrey E. Barrick
jbarrick@cm.utexas.edu
[1] Department of Molecular Biosciences, Center for Systems and Synthetic Biology, The University of Texas at Austin, Austin, Texas, USA
[2] Department of Integrative Biology, The University of Texas at Austin, Austin, Texas, USA

and take over a population [5–7]. Burden has limited the performance and functional lifetimes of cells with complex genetic circuits and metabolic engineering at both laboratory and commercial scales [6, 8–11].

Engineered cells must distribute gene expression resources between native and added functions. Depletion of RNA polymerase molecules, nucleotides, ribosomes, tRNAs, amino acids, or other transcription and translation factors by genes in an engineered DNA construct will reduce the expression of a cell's genes and slow its replication. Of these factors, ribosomes are most commonly cited as the limiting resource when proteins are overexpressed in engineered *E. coli* cells [2, 12–16]. As an mRNA is produced from an engineered DNA construct, ribosomes that would otherwise be allocated to make a cell's native proteins needed for replication are instead diverted to produce the exogenous protein, causing gene expression burden.

Because the concentrations of different tRNA species in cells and the rates at which they are charged by aminoacyl-tRNA synthetases vary, codon usage affects the expression of proteins by impacting translational efficiency, the rate at which ribosomes move along an mRNA as they synthesize the encoded protein [17–22]. Codon usage bias in genes can affect the degree to which their production impacts native gene expression and therefore the burden of protein overexpression [2, 23]. It is common for researchers to adopt codon optimization strategies to achieve higher expression levels of an engineered construct by improving its translational efficiency. Common methods include maximizing Fraction of Optimal codons (FOP) [24] or Codon Adaptation Index (CAI) [25] scores. Alternatively, newer global codon harmonization [26] and codon health index (CHI) [22] metrics can be used. The importance of optimizing codon usage to achieve high levels of protein expression is also evident from how deoptimizing codon usage in the major capsid protein of bacteriophage T7 markedly decreases its fitness [27] and from how other bacteriophages encode their own tRNAs to reprogram host codon usage [28]. Similarly, higher yields of recombinant proteins can be achieved in *E. coli* strains engineered to overexpress rare tRNAs [29].

Researchers must have a robust understanding of cellular resources and when they become limiting to engineer cells that efficiently and stably carry out new functions. To improve our understanding of the relationship between codon usage bias and burden from protein overexpression, we simulated and experimentally characterized overexpression of proteins from genes with a range of different codon usage biases in *E. coli*. We found that there is a negative linear relationship between protein production and burden, regardless of the level of codon (de)optimization, and that codon usage biases which deviate from a cell's charged tRNA availability exacerbate the fitness cost of overexpression.

Results

Gene expression burden depends on codon usage in simulations

Prior studies have shown that there is an inverse relationship between exogenous protein overexpression and the growth rate of a bacterial cell [2, 12, 30]. Less is known about how this relationship is influenced by codon usage. Expression of an mRNA encoding an exogenous gene that has less optimal codon usage is expected to be more burdensome to a cell because its translation requires more rare tRNAs, which are a limiting resource [2, 18, 31, 32]. To test this hypothesis, we developed a computational model for protein overexpression using Pinetree [33], a stochastic gene expression simulation framework. Pinetree models translation by explicitly tracking ribosomes as they move along mRNAs. We updated Pinetree to also simulate dynamic tRNA recycling, in which charged tRNAs are depleted by translation of their cognate codon and then replenished by aminoacyl-tRNA synthetase reactions. These simulations also take into account how tRNA abundance affects the kinetics of translating ribosomes. The rate of decoding a codon and translocating to the next codon is proportional to the concentration of the relevant charged tRNA species.

In the simulations, we used a mixture of two classes of tRNAs, corresponding to optimal and non-optimal codons. We set up simulations so that the concentration of one of the tRNAs is greater than the other, which makes one codon optimal in terms of codon usage bias. Specifically, one tRNA constituted 70% of the whole tRNA pool while the other tRNA constituted the remaining 30%. Growth of cells is proportional to the production of a generic stand-in for all cellular protein (CP), which is encoded by an mRNA that has codon usage biased toward the optimal codon. Then, we add an mRNA encoding an exogenous protein with its own codon usage bias. We simulated different levels of overexpression of the exogenous protein by varying the rate of translation initiation on its mRNA. For the overexpressed protein (OEP), we simulated five different mRNA constructs, with codon usage biases ranging from 10% to 90% optimal. The fitness burden the simulated cell experiences from OEP production is equal to how much its presence decreases expression of CP relative to its baseline level when the OEP mRNA is not present.

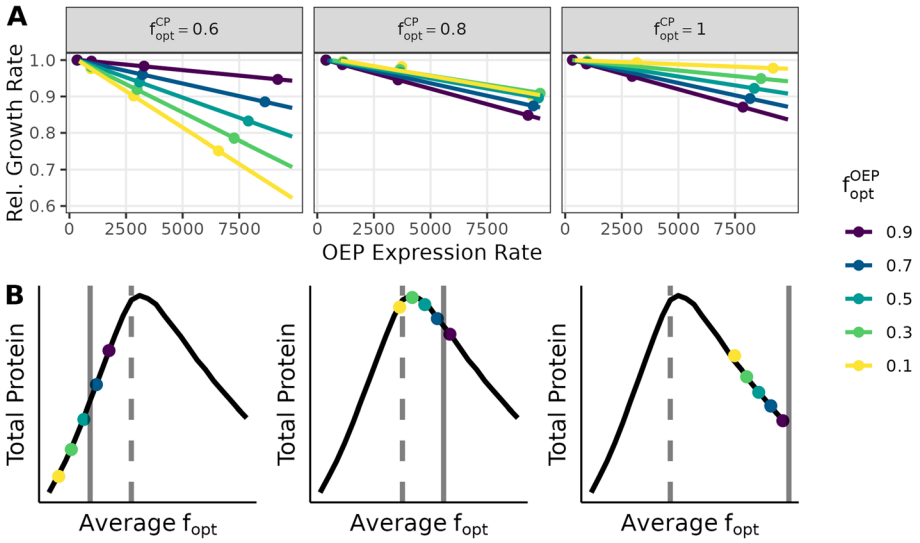

Fig. 1 Simulation of engineered protein overexpression. **A** Simulated cell growth rate versus overexpressed protein (OEP) production rate for different fractions of optimal codon usage in the OEP (f_{opt}^{OEP}) and cellular protein (CP) (f_{opt}^{CP}). In the model, cell growth rates are the amount of CP produced over the duration of the simulation. To get the relative cell growth rate, we divide by the baseline cell growth rate when OEP is not present. OEP expression rate is the number of molecules of protein produced in a simulation divided by its duration. **B** Total protein (CP + OEP) versus average codon usage (weighted average of OEP codon usage and CP codon usage). The colored points in each plot correspond to the simulations shown in panel **A**, when OEP expression is maximized (highest OEP expression rate, points furthest to the right). The solid vertical line indicates f_{opt}^{CP}, and the dashed vertical line indicates the fraction of optimal tRNA (0.7 in all cases)

The simulations captured the expected inverse relationship between exogenous protein expression and the cell growth rate (Fig. 1A). When we made cell codon usage 60% optimal ($f_{opt}^{CP} = 0.6$), burden increased as expected when we decreased f_{opt}^{OEP} (Fig. 1A, left panel). We also ran simulations with higher levels of optimal codon usage for CP, specifically 80% and 100% optimal codon usage. Unexpectedly, we observed qualitatively different simulation behavior when we increased f_{opt}^{CP} in these cases. In the case when CP codon usage was 100% optimal (Fig. 1A, right panel), there was a reversal in the order of the lines with expression of the most optimized OEP construct paradoxically creating the greatest fitness burden.

This reversal can be explained by the observation that in the simulations, 100% optimal codon usage does not always maximize gene expression (Figs. 1B and 2). Specifically, when tRNA pools are partially but not completely charged, protein expression is maximized when the average f_{opt} of the combined set of all transcripts (for CP and OEP together) approximately matches the fraction of optimal tRNAs. (In practice, the exact location of the optimum also depends on the tRNA charging rate, see Fig. 2.) In this regime, the extent to which an overexpressed exogenous gene burdens the cell depends on the direction that expressing the OEP moves the system along the curve relative to its optimum. For example, in simulations where f_{opt}^{CP} is 60% (which is slightly less than the f_{opt} that maximizes expression) overexpressing OEP transcripts that use fewer optimal codons than CP transcripts lowers the average f_{opt}, moving the system away from the maximum point (Fig. 1B, left panel). As a consequence, the fraction of ribosomes sequestered on OEP transcripts increases (Fig. S1), and the cell is less able to produce its own protein. In contrast, when the codon usage of CP is 100% optimal, expressing deoptimized OEP moves the system back in the direction of the optimum (Fig. 1B, right panel). Overall, our simulations predict that burden from protein overexpression depends not only on an exogenous gene's codon usage but also on the extent to which a cell's codon usage matches tRNA abundances.

Gene expression burden depends on codon usage in *E. coli*

We tested the expectations from the model by overexpressing fluorescent proteins (FPs) modified to have different levels of codon usage bias in *E. coli*. We tested both sfGFP and mCherry2 variants. These FPs are derived from evolutionarily distant species [34] and share just 28% amino acid identity. Each FP was synthesized as a set of coding sequence (CDS) variants that used 10%, 25%, 50%, 75%, or 90% optimal codons as defined by Zhou et al. [35]. Highly expressed proteins in *E. coli* have 64% optimal codons by this metric [35, 36] (Fig. S2). Each FP variant was cloned into a plasmid where its expression was controlled by an inducible T7 RNA polymerase promoter and one of five ribosome binding site (RBS)

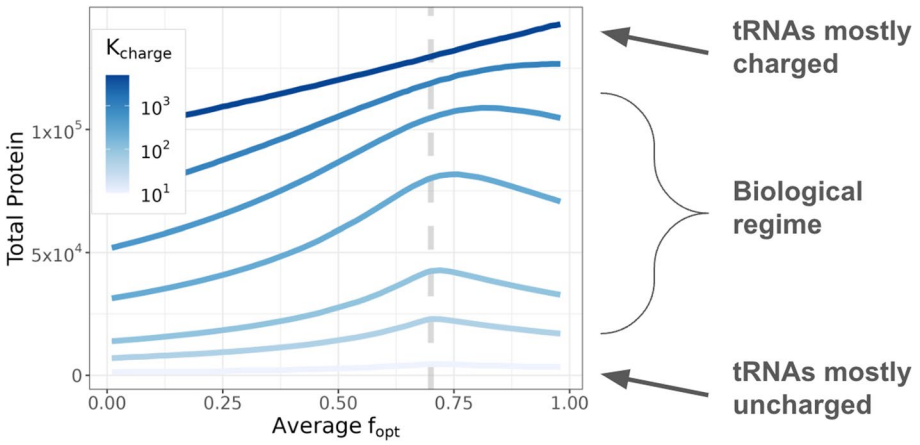

Fig. 2 Relationship between total protein abundance and average f_{opt} in simulations. The dashed vertical line is the optimal tRNA fraction (f_{opt}^{tRNA}), which is 0.7 in this case. When the tRNA charging rate (K_{charge}) is moderate and tRNA pools are partially charged (biological regime) protein expression is maximized when $f_{opt} \approx f_{opt}^{tRNA}$. For very high tRNA charging rates, the amount of total protein produced by a cell is maximized when only optimal codons are used, but the required charging rates are much higher than is realistic for actual cells. K_{charge} rates are in units of s^{-1}

sequences predicted to have a range of translation initiation rates (Fig. 3A). We used these constructs to assess how the relationship between protein overexpression and *E. coli* growth rate depended on codon usage, taking changes in fluorescence after induction during exponential growth as a measure of OEP production.

Codon usage altered the relationship between how much the growth rate of *E. coli* cells slowed as they produced more sfGFP (Fig. 3B). More optimal codon usage increased the highest rate of sfGFP production achieved by a CDS variant. The 10%, 25%, 50%, and 75% optimal sfGFP constructs had maximum expression rates of 0.53, 0.73, 0.89, and 1.04 relative to the maximum expression observed for the 90% optimized variants. We also observed more burden for sfGFP variants with fewer optimal codons even though they did not achieve as high expression. CDS variants engineered with 90%, 75%, or 50% optimal codons had the smallest reductions in growth rate as the rate of protein production increased (Table S1). They reduced the growth rate by 24%, 29%, and 36%, respectively, at the highest levels of expression. Deoptimizing sfGFP to 25% or 10% optimal codons resulted in a marked increase in burden per protein produced, reducing population growth rate by 54% or 61% at the highest levels of expression. These trends can be summarized across all constructs in terms of the negative slope for the linear relationship between the two rates, which we refer to as the overexpression burden cost because it represents how much the host's growth rate decreases for a given increase in the amount of the exogenous protein that it expresses (Fig. 3C).

The mCherry2 CDSs showed a more complex trend in how overexpression burden cost varied with respect to codon usage (Fig. 3D, E). Constructs engineered with 75% and 50% optimal codons performed similarly. mCherry2 with 75% optimal codons reduced the growth rate by 22% at maximum expression, while mCherry2 at 50% optimal codons reduced the growth rate by 30% (Table S1). We observed 28% higher maximum expression with 50% optimized mCherry2 compared to 75% optimized mCherry2. mCherry2 with 25% and 10% optimal codons had more extreme impacts on *E. coli* growth rates as expression level increased. These two constructs reduced growth rates approximately 64% and 71% at maximum expression. The overexpression burden cost for the 90% optimized CDS most resembled that of the 50% optimized CDS, although the difference between the overexpression burden costs of the 75% optimized and 90% optimized CDSs were not statistically significant ($p = 0.16$, likelihood ratio test). Still, these observations could indicate that the 90% optimized mCherry2 CDS is in the overoptimized regime that is predicted to exist in the simulations.

We next tested whether deoptimizing a single codon would lead to similar trends (Fig. 4). We constructed plasmids with variants of the 50% optimized sfGFP CDS that had all of their codons for one amino acid replaced with a single rare codon for that amino acid (Table 1). We performed this targeted deoptimization for four rare codons corresponding to three amino acids (Glu, Gly, and Leu) that appear many times in the sequence of GFP. The Glu codon GAG, which has a 32% genomic frequency and is the only alternative codon to GAA for Glu, had just a 35% reduction in growth rate

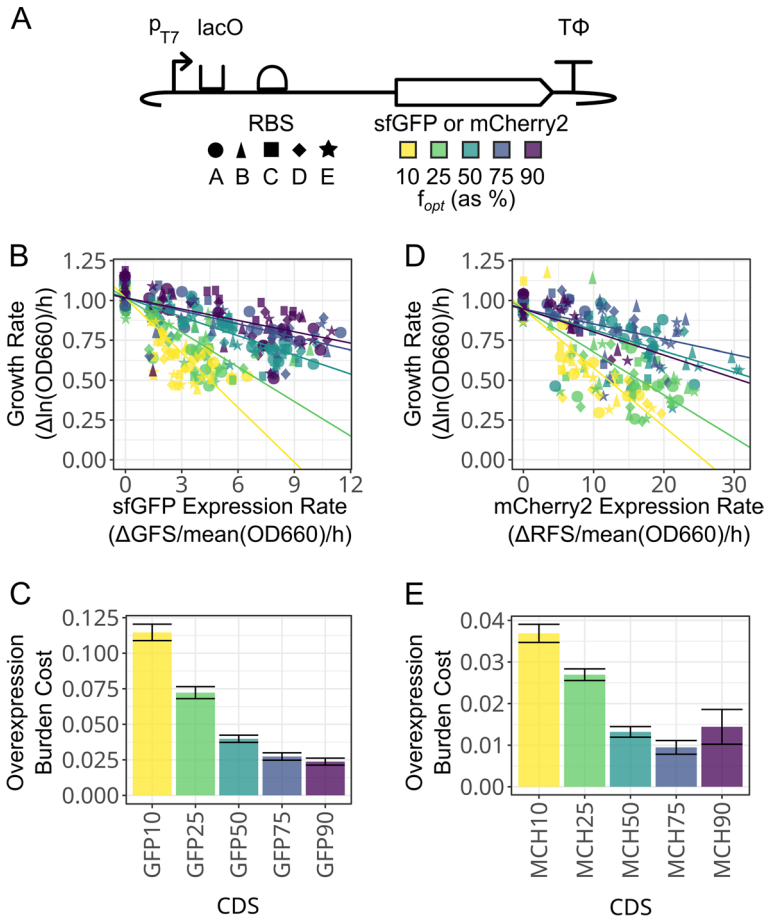

Fig. 3 Effects of changes in codon usage on the relationship between fluorescent protein overexpression and *E. coli* growth rate. **A** Fluorescent protein constructs tested. Five different ribosome binding site (RBS) sequences were combined with five sfGFP and five mCherry2 coding sequence (CDS) variants with different percent codon optimization. The RBS and CDS shape and color legends are used in the other panels. **B** sfGFP construct rates of protein expression versus *E. coli* growth rates. Each line represents a linear regression for the respective CDS variant fit to a common *y*-intercept for no sfGFP expression. Abbreviations: GFS, green fluorescence signal; OD660, optical density at 660 nm. OD and fluorescence readings for each assay are shown in Figs. S3 and S4. **C** Overexpression burden cost for sfGFP constructs (decrease in growth rate per unit increase in the rate of protein expression) determined from the slopes of the linear regressions. Error bars are the standard error of the fits. **D** mCherry2 construct rates of protein expression versus *E. coli* growth rates. Details are as in panel **B**. Abbreviation: RFS, red fluorescence signal. OD and fluorescence readings for each assay are shown in Figs. S5 and S6. **E** Overexpression burden cost for mCherry2 constructs. Details are as in panel C

at maximum protein expression. This is comparable to the 37% reduction observed for the 50% optimized GFP CDS used as a baseline. Deoptimizations targeting the rarer Gly GGG, Gly GGA, and Leu CTA codons all resulted in substantial burden. They reduced growth rates at maximum expression by 63%, 61%, and 63% respectively. These results show that creating a CDS that depletes a single rare tRNA for one amino acid can also lead to high overexpression burden.

Although sfGFP was evolved for improved folding and mCherry2 was evolved for lower cytotoxicity [37, 38], misfolded and potentially toxic aggregates could arise and their levels could vary among our constructs due to how codon usage and mRNA structures alter the speed of translation and affect co-translational folding [20, 39, 40]. Misfolding could induce cell stress responses that impact growth or invalidate the linear relationship between fluorescence and FP gene expression [41]. To assess this possibility, we performed a burden assay on a subset of our constructs and measured soluble and insoluble protein amounts in crude cell lysates. The ratio of purified soluble and insoluble protein was consistent across these constructs (Fig. S9A, $p = 0.43$, ANOVA). The insoluble protein fractions also appeared to have low and consistent amounts of each FP when they were separated on SDS PAGE gels (Fig. S9B, C, D). Therefore, differences in FP misfolding or aggregation are unlikely to have complicated our results.

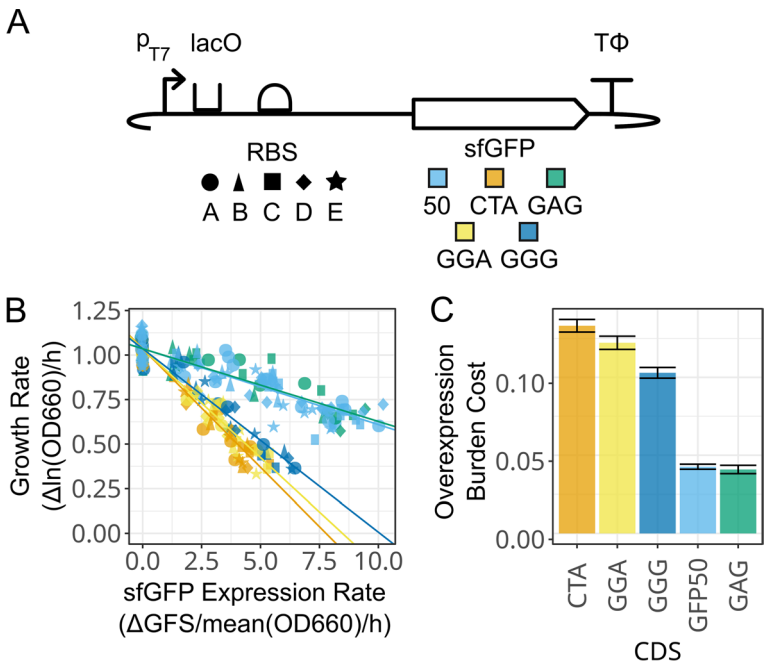

Fig. 4 Effects of replacing codons encoding a single amino acid with a rare codon on the relationship between fluorescent protein overexpression and *E. coli* growth rate. **A** Fluorescent protein constructs tested. Five different RBS sequences were combined with the 50% optimized sfGFP CDS and four variants of this CDS shown in Table 1. The RBS and CDS shape and color legends are used in the other panels. **B** Rates of protein expression versus *E. coli* growth rates. Each line represents a linear regression for the respective CDS variant fit to a common *y*-intercept for no sfGFP expression. OD and fluorescence readings for each assay are shown in Figs. S7 and S8. **C** Overexpression burden cost for sfGFP constructs (decrease in growth rate per increase in the rate of protein expression) determined from the slopes of the linear regressions. Error bars are +/- the standard error of the fits

Discussion

Bacteria such as *E. coli* are commonly engineered to overexpress recombinant proteins for purification and characterization, enzymes to assimilate feedstocks and synthesize bioproducts, and biosensors and regulatory proteins to control their behavior. Here, we used stochastic simulations of gene expression to examine how codon usage influences the relationship between the rate at which a cell expresses a protein of interest and its growth rate. As expected, our simulations predict that replacing rare codons with optimal codons will, in many cases, increase protein expression and reduce the burden on a cell from producing a given amount of protein. Specifically, this is expected to be the case when a sequence begins with codon usage that is poorly matched to the cell's preferred codon usage. However, our simulations also predict that a regime exists in which overusing optimal codons to express the protein of interest can make them become limiting and exacerbate the amount of burden from a given amount of protein production.

To test these predictions, we examined fluorescent protein genes with a range of codon usage biases expressed at different levels in *E. coli*. We found that increasing codon optimization generally led to more fluorescent protein production at less cost to the cell. We also observed elevated genetic instability at both codon usage extremes for mCherry2, including for an

Table 1 Variants of the 50% optimized sfGFP CDS with all codons for a single amino acid replaced by one rare target codon

Amino acid	Rare target codon	Genomic frequency	Optimal to target replacements	Rare to target replacements	CDS f_{opt}
Glu	GAG	32%	12	0	47.2%
Gly	GGA	13%	18	0	44.8%
Gly	GGG	15%	18	2	44.8%
Leu	CTA	3%	16	2	45.8%

mCherry2 gene that seemed to exhibit signs of overoptimization. Differences in the sensitivity of the system to changes in codon usage bias for mCherry2 versus sfGFP are likely due to differences in the amino acid composition of these proteins and therefore which specific tRNA species they deplete. We further showed that deoptimizing a coding sequence by including many copies of a single rare codon for one amino acid could have as severe an effect on gene expression and burden as broadly deoptimizing its sequence to include rare codons for many amino acids.

Changing codon usage can alter other aspects of gene expression and burden that are not accounted for in our simulations. We did not alter the first 15 codons of our fluorescent protein genes to minimize well-known effects of RNA structure around the ribosome binding site and start codon on translation initiation rates [32]. RNA structures formed later in their reading frames may be highly variable between our constructs that have different codon usage. These structures could slow translation, leading to less expression and more burden [42–44]. Either of these effects of mRNA structures could explain some of the variation in expression between the constructs we tested.

Several studies have explored relationships between burden, gene expression, and codon usage in *E. coli*. One examined the effects of deoptimizing codons at the 3' end of the *vioB* gene [2]. They predicted that adding rare codons would make gene expression burden more severe because slowing translation at this location would sequester more ribosomes on this mRNA, and they found this effect. Another study that compared several optimization schemes based on the codon adaptation index (CAI) also had difficulty expressing maximally codon optimized constructs [26]. A codon health index (CHI) was developed as an alternative to CAI by measuring how much use of different codons competed with native cellular proteins, as a way of coding for high expression while minimizing burden on the host cell [22]. *E. coli* cell-free transcription-translation (TX-TL) systems have been used to examine extreme codon usage while avoiding complications from evolutionary instability [22, 45].

We used Pinetree [33] to simulate dynamic charging and depletion of tRNA pools while ribosomes bind to and translate mRNAs with single-codon resolution. Pinetree has been used to simulate other bacterial gene expression processes, including transcription of mRNAs from a genome, co-transcriptional translation of mRNAs, and mRNA degradation [46, 47]. Either Pinetree or a variety of similar models that others have used to examine codon usage [2, 22, 45, 48, 49] could be extended to improve our understanding of how different bioengineering design choices affect burden. Ultimately, these simulations could refine the predictions made by design tools, such as the Operon Calculator [50] and CryptKeeper [51], so researchers can maximize protein overexpression while minimizing burden on engineered cells and reducing the incidence of evolutionary failure.

Methods

Gene expression simulation framework

For the computational modeling, we used the Pinetree [33] gene expression simulation framework, which implements a version of the Gillespie Stochastic Simulation algorithm [52] for modeling transcription and translation dynamics at the nucleotide or codon level. In the previously published version of Pinetree, tRNAs were not explicitly simulated, and per-codon translation rates could not change dynamically during a simulation. We modified Pinetree to explicitly account for charged tRNA availability and tRNA re-charging. The updated version of Pinetree is also different from the prior version in that the per-codon translation rates depend on the availability of charged tRNA pools, which can change dynamically depending on codon usage. In the new model, pools of charged tRNAs can go to zero, if codon usage is very biased and/or the tRNA charging rate is very low. Ribosomes occupying codons with depleted tRNA pools will stall until newly charged tRNAs become available.

Optimal codon usage model

To simulate optimal versus non-optimal codon usage in Pinetree, we used a simplified model where the total tRNA population consists of just two species—an optimal tRNA that comprises the the majority of the tRNA pool (greater than 50%), and a non-optimal tRNA that makes up the remainder of the tRNA pool. In this model, an optimal codon is a codon that corresponds to an optimal tRNA, and all transcripts are comprised of some combination of optimal and non-optimal codons. The codon-specific translation propensity (similar to the rate in an ODE model) for optimal codon translation is:

$$R_b^{\text{opt}} T_c^{\text{opt}} K_{\text{speed}}\,, \tag{1}$$

where R_b^{opt} is the number of ribosomes bound to an optimal codon, T_c^{opt} is the number of charged optimal tRNAs, and K_{speed} is the baseline speed of the ribosome in codons per second. Similarly, the expression for the non-optimal codon translation propensity is:

$$R_b^{\text{nopt}} T_c^{\text{nopt}} K_{\text{speed}}\,, \tag{2}$$

Table 2 Parameters for overexpression simulations

Parameter	Value(s)	Units
K_{speed}	2	codons/s
K_{charge}	100	s^{-1}
K_{bind} (CP RBS)	1×10^7	$M^{-1}s^{-1}$
K_{bind} (OEP RBS)	1×10^6–1×10^8	$M^{-1}s^{-1}$
f_{opt}^{CP}	0.6, 0.8, 1.0	-
f_{opt}^{OEP}	0.1–0.9	-
f_{opt}^{tRNA}	0.7	-

where R_b^{nopt} and T_c^{nopt} are the number of ribosomes bound to non-optimal codons and the number of non-optimal charged tRNAs, respectively. In general, R_b^{opt}, R_b^{nopt}, T_c^{opt}, and T_c^{nopt} are dynamic values that can change during the simulation. In this study, these rapidly reach a steady-state within the first 10 seconds of the simulation. The simulation-specific steady-states depend on the f_{opt} values for OEP mRNA and CP mRNA, and the reaction rate constants for tRNA charging, ribosome binding, and ribosome elongation.

Simulating exogenous protein overexpression

To simulate overexpression of an exogenous protein in a bacterial cell, we set up simulations with 100 identical copies of a 3000-bp mRNA transcript that represents cellular protein (CP). We then added 20 mRNA transcripts of 300-bp each that encode the overexpressed protein (OEP). We made the simplifying assumption that transcript abundances are at a steady-state and that transcription dynamics can be ignored. Each simulation begins with 500 ribosomes and 2500 total tRNAs, which are initially unbound and uncharged but quickly reach an equilibrium, where the average numbers of bound ribosomes and charged tRNAs no longer change. For the three major species in the simulation (transcripts, ribosomes, and tRNAs), we used initial total counts that are approximately 1/100 th the actual abundances for each species in an *E. coli* cell during exponential growth [53, 54]. Cell growth rates and OEP production rates were calculated by taking CP and OEP abundances after 100 seconds of simulated time. Because simulations are non-deterministic, for every individual simulation we ran 3 replicates and used protein abundances averaged over these replicates. To get steady-state charged tRNA and steady-state bound ribosomes amounts, we took an average over the last 50 seconds of the simulation run, after it had reached equilibrium.

Calibration of simulation parameters

The major free parameters in these simulations are the rate constants for tRNA charging (K_{charge}), ribosome s(K_{speed}). To calibrate these parameters, we manually fit ribosome and tRNA steady-states in the baseline simulation (when OEP is not present) to empirical measurements. In exponentially growing *E. coli* cells, at least 80% of the total ribosome content is bound to mRNA [55], and 75% of the tRNA pool is charged with an amino acid [56]. Using these values as our target, we set K_{bind} to 1×10^7 $M^{-1}s^{-1}$ (which we had used in prior simulations [46]) and then chose values for K_{charge} and K_{speed} that achieved close to the target steady-state quantities. Figure S10 shows the entire grid search over this parameter space. Finally, for f_{opt}^{tRNA}, we chose a value of 0.7, which means that 70% of the total tRNA pool in the simulation is optimal tRNA and 30% is non-optimal tRNA. (Note that the fractions of *charged* tRNAs can and will differ from these values, depending on the codon usage of the CP and OEP and various simulation parameters.) A summary of all simulation parameters and their values is provide in Table 2.

Plasmid design and assembly

DNA sequences encoding mCherry2 and sfGFP CDSs with different codon usage biases were generated using a Python script that randomly replaced codons between optimal and nonoptimal possibilities until the desired level of optimization was reached. The first 15 codons were left unchanged to reduce the chances of substantially altering translation initiation rates [32]. Additionally, we mutated mCherry2 to eliminate an alternative start codon [57]. Optimal codon classifications were from a previous analysis of highly expressed *E. coli* genes [35]. For sfGFP genes in which we deoptimized all codons for one amino acid, 2–3 codons for other amino acids had to be mutated from optimal to rare codons to make the sequence compatible with DNA synthesis. Primers and plasmids are shown in Tables S2 & S3. The full coding sequences tested are provided in Tables S4, S5, & S6.

We initially cloned these CDS constructs into plasmid pET21 and performed experiments using variable concentrations of IPTG to achieve different expression levels, but this resulted in differences in the timing of induction that complicated data analysis [58]. Therefore, we designed plasmids in which we cloned new RBS sequences predicted to have 0.25 (A), 0.5 (B), 2 (C), or 4 (D) times the strength of the pET21 vector RBS (E). Alternative ribosome binding sites were generated *in silico* by introducing random mutations and predicting translation initiation rates using OSTIR [59] until they matched the desired strength in the context of a complete sfGFP expression plasmid.

CDS variants were ordered as gBlocks (Integrated DNA Technologies). Golden Gate assembly overhangs and alternative RBS sequences were either included in the

gBlocks or added in PCR primers. Plasmid pET21b(+) was domesticated to remove three BsmBI cut sites using PaqCI Golden Gate assembly (GGA). Parts were first assembled into an insulated part plasmid [60] using BsmBI GGA and chemically transformed into *E. coli* NEB 5-alpha cells (New England Biolabs). Part plasmids and the pET21b(+) backbone were then assembled into complete expression constructs using PaqCI GGA, chemically transformed into NEB 5-alpha, and subcloned into *E. coli* Tuner(DE3) cells (Novagen). Plasmids were confirmed to have the designed sequence and to be monomeric using nanopore sequencing (Plasmidsaurus).

Fluorescent protein overexpression burden assays

E. coli strains with FP overexpression plasmids were cultured in LB supplemented with 100 μg/ml carbenicillin (Crb). Strains were struck from −80°C glycerol stocks onto LB + Crb plates and allowed to grow overnight at 30 °C or 37 °C. The next day, single colonies were picked at random and allowed to grow overnight in 5 mL LB + Crb at 30 °C or 37 °C with orbital shaking at 200 RPM over a 1-inch diameter. 10 μL of overnight culture was added to 190 μL of fresh LB + Crb in a 96-well microplate and allowed to precondition in a Tecan Infinite Pro M200 microplate reader at 37 °C with orbital shaking over a 1-mm radius at 432 RPM for 11 min 40 sec per 15-min measurement cycle for a total of 2 to 2.5 hr. Then, cultures were normalized to an OD660 reading of approximately 0.2 and induced by adding 0.19 mM isopropyl *β*-D-1-thiogalactopyranoside (IPTG) using an automated liquid handler or 0 mM to measure uninduced growth rates. This inducer concentration was optimized by conducting preliminary tests of the 50% optimal GFP constructs with 0.07 mM to 0.25 mM IPTG and selecting a concentration that produced a high dynamic range of growth rates. To control for position effects in the plate, cells with different constructs were randomly distributed across wells, excluding the perimeter when possible. Cells were then induced and immediately returned to the microplate reader. OD660 and fluorescence measurements were taken every 15 minutes. sfGFP was measured at 480 nm excitation and 525 nm emission wavelengths. mCherry2 was measured at 590 nm excitation and 645 nm emission wavelengths.

Following an approach used in prior studies [2, 61], *E. coli* growth rate and FP expression rates were analyzed in 1-hour sliding windows. Growth rate was defined as $\frac{\ln OD660_{t2} - \ln OD660_{t1}}{(t2-t1)}$. FP expression rate was defined as $\frac{FS_{t2} - FS_{t1}}{(t2-t1) * mean(OD660_{t2-t1})}$, where FS is the fluorescent signal from sfGFP or mCherry2. The starting time for windows used to fit growth rates was 1 h 15 min post-induction for both GFP and mCherry2. FP expression rates were fit from time windows beginning 1 h 30 min post-induction for sfGFP variants and 2 h 15 min post-induction for mCherry2 variants. These times were selected to correspond to when there were the largest growth rate differences among constructs and maximal FP production rates. The longer delay in measuring mCherry2 expression reflects its slower maturation time.

For analyzing the relationship between growth and expression rates, linear regression was performed with a common *y*-intercept (zero expression) within each group of CDS variants (percent optimized sfGFP, percent optimized mCherry2, or single-codon deoptimized sfGFP). Some variation in FP expression with a proportional change in growth rate among replicate assays of the same construct appears to be due to differences in plasmid copy number that arise due to random segregation and then selection favoring cells with reduced copy number after induction. Wells for a given RBS and FP variant that exhibited less than 10% of the maximal normalized fluorescence observed for that construct across all wells were dropped from the analysis since they were uninformative for fitting slopes. The mean of the top 10% observed expression values for each CDS variant was used for reporting growth rates at maximal expression. The final parameters fit for each construct are provided in Table S1.

Protein analysis

Strains were struck from −80°C glycerol stocks onto LB + Crb plates. Each strain was picked in triplicate and allowed to grow overnight in 5 mL LB + Crb. The next morning, 1 mL of overnight culture was reserved at 4°C. Multiple wells of each sample were assayed following the burden assay protocol with slight deviations. The shaking time per kinetic cycle was reduced to 675 seconds to account for the slightly increased fluorescence measuring time and induction was performed via manual pipetting. After the burden assay, wells with same construct were repooled. Then, all samples, including the reserved overnight cultures, were pelleted via centrifugation at 3000 × g for 5 min and frozen at −80°C until analysis.

Cells were lysed using B-PER Bacterial Protein Extraction Reagent (Thermo Fisher Scientific). Cells were allowed to thaw at 4°C and refrozen at 80°C for 30 min for a total of 3 thaw periods. Then, cell pellets were resuspended in 80 μL lysis buffer and allowed to incubate at room temperature for 30 min. Lysed cells were centrifuged for 10 minutes at 15,000 × g to separate soluble and insoluble material. The supernatant was reserved for downstream analysis as the soluble fraction. The insoluble fraction was washed with 80 μL lysis buffer and recentrifuged twice, and then resuspended once more in 80 μL lysis buffer.

For measuring the total protein content of the soluble and insoluble fractions, 20 μL lysis buffer with 10% sodium dodecyl sulfate (SDS) was added to each 80 μL sample. Protein concentrations were measured using a

Pierce BCA Protein Assay Kit (Thermo Fisher Scientific) using an automated liquid handler, following the manufacturer's microplate protocol with a sample:reagent ratio of 1:10. For examining proteins in different fractions, approximately 1 μg of total protein was loaded on a NuPAGE 4–12% Bis-Tris polyacrylamide gel in MES SDS buffer. After electrophoresis, gels were stained using a short silver nitrate staining protocol [62].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13036-025-00521-z.

Supplementary Material 1

Acknowledgements
The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high-performance computing resources.

Authors'contributions
Cameron T. Roots and Alexis M. Hill contributed equally. Cameron T. Roots: Conceptualization; Software; Formal Analysis; Investigation; Visualization; Methodology; Writing—original draft; Writing—review and editing. Alexis M. Hill: Software; Formal Analysis; Investigation; Visualization; Methodology; Writing—original draft; Writing—review and editing. Claus O. Wilke: Conceptualization; Formal Analysis; Supervision; Methodology; Funding acquisition; Writing— review and editing. Jeffrey E. Barrick: Conceptualization; Formal Analysis; Supervision; Funding acquisition; Writing—original draft; Writing—review and editing.

Funding
This work was supported by the National Institutes of Health (R01GM088344 to COW and JEB) and the National Science Foundation (MCB-2123996 to JEB). COW also acknowledges support from the Blumberg Centennial Professorship in Molecular Evolution. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability
Code and source data for the simulations and experiments in this study are available in Zenodo (https://doi.org/10.5281/zenodo.14232327).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 6 December 2024 Accepted: 9 May 2025
Published online: 23 May 2025

References

1. Nevozhay D, Adams RM, Van Itallie E, Bennett MR, Balázsi G. Mapping the environmental fitness landscape of a synthetic gene circuit. PLoS Comput Biol. 2012;8(4):e1002480. https://doi.org/10.1371/journal.pcbi.1002480.
2. Ceroni F, Algar R, Stan GB, Ellis T. Quantifying cellular capacity identifies gene expression designs with reduced burden. Nat Methods. 2015;12:415–8. https://doi.org/10.1038/nmeth.3339.
3. Nikolados EM, Weiße AY, Ceroni F, Oyarzún DA. Growth defects and loss-of-function in synthetic gene circuits. ACS Synth Biol. 2019;8(6):1231–40. https://doi.org/10.1021/acssynbio.8b00531.
4. Santos-Navarro FN, Vignoni A, Boada Y, Picó J. RBS and promoter strengths determine the cell-growth-dependent protein mass fractions and their optimal synthesis rates. ACS Synth Biol. 2021;10(12):3290–303. https://doi.org/10.1021/acssynbio.1c00131.
5. Sleight SC, Bartley BA, Lieviant JA, Sauro HM. Designing and engineering evolutionary robust genetic circuits. J Biol Eng. 2010;4:12.
6. Rugbjerg P, Myling-Petersen N, Porse A, Sarup-Lytzen K, Sommer MOA. Diverse genetic error modes constrain large-scale bio-based production. Nat Commun. 2018;9(1):787. https://doi.org/10.1038/s41467-018-03232-w.
7. Radde N, Mortensen GA, Bhat D, Shah S, Clements JJ, Leonard SP, et al. Measuring the burden of hundreds of BioBricks defines an evolutionary limit on constructability in synthetic biology. Nat Commun. 2024;15(1):6242. https://doi.org/10.1038/s41467-024-50639-9.
8. Sandoval CM, Ayson M, Moss N, Lieu B, Jackson P, Gaucher SP, et al. Use of pantothenate as a metabolic switch increases the genetic stability of farnesene producing Saccharomyces cerevisiae. Metab Eng. 2014;25:215–26. https://doi.org/10.1016/j.ymben.2014.07.006.
9. Burgard A, Burk MJ, Osterhout R, Van Dien S, Yim H. Development of a commercial scale process for production of 1,4-butanediol from sugar. Curr Opin Biotechnol. 2016;42:118–25. https://doi.org/10.1016/j.copbio.2016.04.016.
10. Zhang S, Voigt CA. Engineered dCas9 with reduced toxicity in bacteria: implications for genetic circuit design. Nucleic Acids Res. 2018;46(20):11115–25. https://doi.org/10.1093/nar/gky884.
11. Rugbjerg P, Dyerberg ASB, Quainoo S, Munck C, Sommer MOA. Short and long-read ultra-deep sequencing profiles emerging heterogeneity across five platform Escherichia coli strains. Metab Eng. 2021;65:197–206. https://doi.org/10.1016/j.ymben.2020.11.006.
12. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of cell growth and gene expression: origins and consequences. Science. 2010;330(6007):1099–102. https://doi.org/10.1126/science.1192588.
13. Qian Y, Huang HH, Jiménez JI, Vecchio DD. Resource compitition shapes the response of genetic circuits. ACS Synth Biol. 2017;6:1263–72. https://doi.org/10.1021/acssynbio.6b00361.
14. Dourado H, Lercher MJ. An analytical theory of balanced cellular growth. Nat Commun. 2020;11(1):1226. https://doi.org/10.1038/s41467-020-14751-w.
15. Dai X, Zhu M, Warren M, Balakrishnan R, Patsalo V, Okano H, et al. Reduction of translating ribosomes enables Escherichia coli to maintain elongation rates during slow growth. Nat Microbiol. 2016;2(2):16231. https://doi.org/10.1038/nmicrobiol.2016.231.
16. Weiße AY, Oyarzún DA, Danos V, Swain PS. Mechanistic links between cellular trade-offs, gene expression, and growth. Proc Natl Acad Sci USA. 2015;112(9):E1038–47. https://doi.org/10.1073/pnas.1416533112.
17. Gingold H, Pilpel Y. Determinants of translation efficiency and accuracy. Mol Syst Biol. 2011;7(1):481. https://doi.org/10.1038/msb.2011.14.
18. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2011;12(1):32–42. https://doi.org/10.1038/nrg2899.
19. Quax TEF, Claassens NJ, Söll D, van der Oost J. Codon bias as a means to fine-tune gene expression. Mol Cell. 2015;59(2):149–61. https://doi.org/10.1016/j.molcel.2015.05.035.
20. Liu Y. A code within the genetic code: codon usage regulates co-translational protein folding. Cell Commun Signal. 2020;18(1):145. https://doi.org/10.1186/s12964-020-00642-6.
21. Schmidt M, Lee N, Zhan C, Roberts JB, Nava AA, Keiser LS, et al. Maximizing heterologous expression of engineered type I polyketide synthases: investigating codon optimization strategies. ACS Synth Biol. 2023;11:3366–80.
22. Love AM, Nair NU. Specific codons control cellular resources and fitness. Sci Adv. 2024;10:eadk3485. https://doi.org/10.1126/sciadv.adk3485.
23. Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM. Codon usage of highly expressed genes affects proteome-wide translation efficiency.

Proc Natl Acad Sci USA. 2018;115:E4940–9. https://doi.org/10.1073/pnas.1719375115.
24. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol. 1981;146(1):1–21. https://doi.org/10.1016/0022-2836(81)90363-6.
25. Sharp PM, Li WH. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987;15(3):1281–95. https://doi.org/10.1093/nar/15.3.1281.
26. Mignon C, Mariano N, Stadthagen G, Lugari A, Lagoutte P, Donnat S, et al. Codon harmonization - going beyond the speed limit for protein expression. FEBS Lett. 2018;592(9):1554–64. https://doi.org/10.1002/1873-3468.13046.
27. Bull JJ, Molineux IJ, Wilke CO. Slow fitness recovery in a codon-modified viral genome. Mol Biol Evol. 2012;29(10):2997–3004. https://doi.org/10.1093/molbev/mss119.
28. Yang JY, Fang W, Miranda/Sanchez F, Brown JM, Kauffman KM, Avevero MC, et al. Degradation of host translational machinery drives tRNA acquisition in viruses. Cell Syst. 2021;12:771–779. https://doi.org/10.1016/j.cels.2021.05.019.
29. Lipinszki Z, Vernyik V, Farago N, Sari T, Puskas LG, Blattner FR, et al. Enhancing the translational capacity of *E. coli* by resolving the codon bias. ACS Synth Biol. 2018;7(11):2656–2664. https://doi.org/10.1021/acssynbio.8b00332.
30. Chure G, Cremer J. An optimal regulation of fluxes dictates microbial growth in and out of steady state. eLife. 2023;12:e84878. https://doi.org/10.7554/elife.84878.
31. Bulmer M. The selection-mutation-drift theory of synonymous codon usage. Genetics. 1991;129(3):897–907. https://doi.org/10.1093/genetics/129.3.897.
32. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in Escherichia coli. Science. 2009;324:255–8. https://doi.org/10.1126/science.1170160.
33. Jack BR, Wilke CO. Pinetree: a step-wise gene expression simulator with codon-specific translation rates. Bioinformatics. 2019;35:4176–8.
34. Matz MV, Fradkov AF, Labas YA, Savitsky AP, Zaraisky AG, Markelov ML, et al. Fluorescent proteins from nonbioluminescent Anthozoa species. Nat Biotechnol. 1999;17(10):969–73. https://doi.org/10.1038/13657.
35. Zhou T, Weems M, Wilke CO. Translationally optimal codons associate with structurally sensitive sites in proteins. Mol Biol Evol. 2009;26(7):1571–80. https://doi.org/10.1093/molbev/msp070.
36. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. Nature. 2004;429(6987):92–6. https://doi.org/10.1038/nature02456.
37. Pédelacq JD, Cabantous S, Tran T, Terwilliger TC, Waldo GS. Engineering and characterization of a superfolder green fluorescent protein. Nat Biotechnol. 2006;24(1):79–88.
38. Shen Y, Chen Y, Wu J, Shaner NC, Campbell RE. Engineering of mCherry variants with long Stokes shift, red-shifted fluorescence, and low cytotoxicity. PLoS ONE. 2017;12(2):e0171257. https://doi.org/10.1371/journal.pone.0171257.
39. Sander IM, Chaney JL, Clark PL. Expanding Anfinsen's principle: contributions of synonymous codon selection to rational protein design. J Am Chem Soc. 2014;136(3):858–61. https://doi.org/10.1021/ja411302m.
40. Quandt EM, Traverse CC, Ochman H. Local genic base composition impacts protein production and cellular fitness. PeerJ. 2018;6:e4286. https://doi.org/10.7717/peerj.4286.
41. Walsh IM, Bowman MA, Soto Santarriaga IF, Anabel R, Clark PL. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. Proc Natl Acad Sci USA. 2020;117(7):3528–34. https://doi.org/10.1073/pnas.1907126117.
42. Wen JD, Lancaster L, Hodges C, Zeri AC, Yoshimura SH, Noller HF, et al. Following translation by single ribosomes one codon at a time. Nature. 2008;452(7187):598–603. https://doi.org/10.1038/nature06716.
43. Bao C, Loerch S, Ling C, Korostelev AA, Grigorieff N, Ermolenko DN. mRNA stem-loops can pause the ribosome by hindering A-site tRNA binding. eLife. 2020;9:e55799. https://doi.org/10.7554/elife.55799.
44. Landick R. Transcriptional pausing as a mediator of bacterial gene regulation. Annu Rev Microbiol. 2021;75(1):291–314. https://doi.org/10.1146/annurev-micro-051721-043826.
45. Borkowski O, Bricio C, Murgiano M, Rothschild-Mancinelli B, Stan GB, Ellis T. Cell-free prediction of protein expression costs for growing cells. Nat Commun. 2018;9(1):1457. https://doi.org/10.1038/s41467-018-03970-x.
46. Jack BR, Boutz DR, Paff ML, Smith BL, Wilke CO. Transcript degradation and codon usage regulate gene expression in a lytic phage. Virus Evol. 2019;5(2):vez055. https://doi.org/10.1093/ve/vez055.
47. Hill AM, Ingle TA, Wilke CO. A computational model for bacteriophage ϕ X174 gene expression. PLoS ONE. 2024;19(10):e0313039. https://doi.org/10.1371/journal.pone.0313039.
48. Reuveni S, Meilijson I, Kupiec M, Ruppin E, Tuller T. Genome-scale analysis of translation elongation with a ribosome flow model. PLoS Comput Biol. 2011;7(9):e1002127. https://doi.org/10.1371/journal.pcbi.1002127.
49. Shallom D, Naiger D, Weiss S, Tuller T. Accelerating whole-cell simulations of mRNA translation using a dedicated hardware. ACS Synth Biol. 2021;10(12):3489–506. https://doi.org/10.1021/acssynbio.1c00415.
50. Tian T, Salis HM. A predictive biophysical model of translational coupling to coordinate and control protein expression in bacterial operons. Nucleic Acids Res. 2015;43(5):7137–51. https://doi.org/10.1093/nar/gkv635.
51. Roots CT, Barrick JE. CryptKeeper: a negative design tool for reducing unintentional gene expression in bacteria. 2024. https://doi.org/10.1101/2024.09.05.611466.
52. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. J Phys Chem. 1977;81:2340–61.
53. Dong H, Nilsson L, Kurland CG. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. J Mol Biol. 1996;260:649–63.
54. Neidhardt FC, Ingraham JL, Low KB, Schaechter M, Magasanik B, Umbarger HE. Escherichia coli and Salmonella typhimurium: cellular and molecular biology. Washington: American Society for Microbiology; 1987.
55. Metelev M, Lundin E, Volkov IL, Gynnå AH, Elf J, Johansson M. Direct measurements of mRNA translation kinetics in living cells. Nat Commun. 2022;13:1852.
56. Evans ME, Clark WC, Zheng G, Pan T. Determination of tRNA aminoacylation levels by high-throughput sequencing. Nucleic Acids Res. 2017;45:e133.
57. Fages-Lartaud M, Tietze L, Elie F, Lale R, Hohmann-Marriott MF. mCherry contains a fluorescent protein isoform that interferes with its reporter function. Front Bioeng Biotechnol. 2022;10:892138. https://doi.org/10.3389/fbioe.2022.892138.
58. Fried VA. A novel mutant of the lac transport system of Escherichia coli. J Mol Biol. 1977;114(4):477–90. https://doi.org/10.1016/0022-2836(77)90173-5.
59. Roots CT, Lukasiewicz A, Barrick JE. 'OSTIR': open source translation initiation rate prediction. J Open Source Softw. 2021;6(64):3362. https://doi.org/10.21105/joss.03362.
60. Lariviere PJ, Zuberi Ashraf AHM, Navarro-Escalante L, Leonard SP, Miller LG, Moran NA, et al. One-step genome engineering in bee gut bacterial symbionts. mBio. 2024;15:e01392–24. https://doi.org/10.1128/mbio.01392-24.
61. Ceroni F, Boo A, Furini S, Gorochowski TE, Borkowski O, Ladak YN, et al. Burden-driven feedback control of gene expression. Nat Methods. 2018;15:387–93. https://doi.org/10.1038/nmeth.4635.
62. Chevallet M, Luche S, Rabilloud T. Silver staining of proteins in polyacrylamide gels. Nat Protoc. 2006;1:1852–8. https://doi.org/10.1038/nprot.2006.288.

Publisher's Note