Noise robustness and metabolic load determine the principles of central dogma regulation

Teresa W. Lo,[1,*] Han Kyou James Choi,[1,*] Dean Huang,[1] and Paul A. Wiggins[1,2,3,†]

[1]*Department of Physics, University of Washington, Seattle, Washington 98195, USA*
[2]*Department of Bioengineering, University of Washington, Seattle, Washington 98195, USA*
[3]*Department of Microbiology, University of Washington, Seattle, Washington 98195, USA*

Protein expression levels optimize cell fitness: Too low of an expression level of essential proteins slows growth by compromising the function of essential processes, whereas the overexpression of proteins slows growth by increasing the metabolic load. This trade-off naïvely predicts that the cell maximizes its fitness by a Goldilocks principle in which cells express just enough protein for function; however, this strategy neglects the significance of the inherent stochasticity of the gene expression process which leads to significant cell-to-cell variation in protein numbers. How does the cell ensure robust growth in the face of the inherent stochasticity of these central dogma processes? To explore the consequences of noise in the expression of hundreds of essential genes, we build a minimal model where the trade-off between metabolic cost and growth robustness can be analyzed analytically. The model predicts that growth-rate maximization leads to a highly-asymmetric cost-benefit analysis which drives the optimal protein expression levels far above what is required on average, with low-expression essential proteins expressed at more than 10-fold what is required for growth in the typical cell. This robustness mechanism naturally explains the surprisingly strong buffering effect observed in essential protein depletion experiments and leads to a second qualitative prediction: there is a lower floor on the transcription level of essential genes corresponding to one message per cell cycle. We show that nearly all essential, but not non-essential, genes obey this limit in three evolutionarily-divergent model organisms: *Escherichia coli*, yeast, and human. The model also predicts that the optimal translation efficiency is roughly proportional to message abundance, predicting the observed relation between proteome fraction and message abundance in eukaryotic cells. This optimal translation efficiency predicts a non-canonical scaling of gene expression noise with protein abundance, which we show is observed in yeast. Together, these results reveal that noise robustness and metabolic load determine the global regulatory principles that govern central dogma function.

INTRODUCTION

What rationale determines the transcription and translation level of a gene in the cell? Both experiment and theory support the idea that gene expression levels maximize cell fitness and that cells can rapidly adapt genetically to new environments [1, 2]. Although it is clear that fitness optimization is the rationale for protein expression levels, the consequences of this optimization on protein expression are still poorly understood. For instance, multiple approaches suggest that protein expression is much higher than one would predict based on protein activity [3–10]. How can these elevated protein levels be reconciled with fitness optimization? One possibility is that growth robustness may explain not only this putative protein overabundance but the relative levels of both transcription and translation. We explore this hypothesis in the paper.

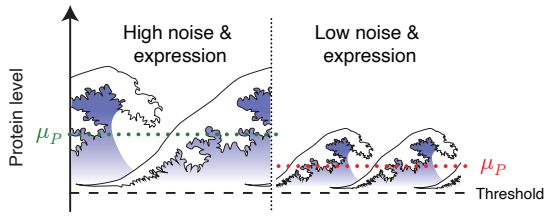

FIG. 1. **Strategies for robustness to noise.** We schematically represent the stochastic protein number as waves. To achieve robust function, the protein levels must be above the threshold. Robustness can be achieved by two distinct strategies: (i) increased expression level and (ii) decreased noise.

Achieving growth robustness is nontrivial since all processes at the cellular scale are stochastic, including gene expression [11]. This biological noise leads to significant cell-to-cell variation in protein numbers, even for essential proteins that are required for growth [12, 13]. How does the cell ensure robust expression of hundreds of distinct essential gene products required for cellular function? Fig. 1 illustrates two qualitative strategies for achieving robustness: (i) increased protein expression levels and (ii) reducing the noise amplitude. Our analysis will demonstrate that the central dogma is regulated to implement both approaches, with strategy (i) leading to protein overabundance and strategy (ii) shaping the relation between transcription and translation.

To study the consequences of growth robustness on the central dogma quantitatively, we propose and analyze a minimal model: the Robustness-Load Trade-Off (RLTO) Model. The model includes three critical components: (i) Protein levels are stochastic and the single-cell growth rate depends upon them, (ii) gene transcrip-

* These two authors contributed equally
† pwiggins@uw.edu

tion and translation generate a metabolic load, and (iii) cell growth is dependent on a large number of essential genes (*i.e.* high multiplicity). Implementing this model required a key theoretical innovation: the analysis of the consequences of noise on a highly-asymmetric fitness landscape. Our novel approach predicts new phenomenology absent from previous models (*e.g.* [14]).

In the RLTO model, growth rate maximization predicts protein overabundance generically and vast overabundance for low-expression essential genes. Protein overabundance explains the paradox of protein expression levels being simultaneously optimal and in excess of what is required for function [9, 15–17]. The model predicts that there is a transcriptional floor of roughly one message transcribed per cell cycle for essential genes. We demonstrate that just such a lower threshold is observed in *Escherichia coli*, yeast, and human. The RLTO model also predicts a central dogma regulatory program that balances transcription and translation, which we call *load balancing*. Load balancing predicts the dependence of gene proteome fraction on message number, as well as the global regulatory response of cells to changes in the metabolic cost of transcription. Furthermore, load balancing predicts that noise should have a non-canonical scaling with protein abundance. We demonstrate that the predicted scaling is observed in yeast. Taken together, these results reveal that noise robustness and metabolic load fundamentally shape the function of the central dogma processes and that global function is quantitatively described by a few simple emergent principles.

RESULTS

Defining the RLTO Model. We start by developing a minimal quantitative model for cell fitness (*i.e.* growth rate), which includes the stochasticity of gene expression, the dependence of growth on essential proteins, as well as the metabolic load of gene expression. We will assume the numbers of proteins for gene i, N_{pi}, are gamma-distributed independent random variables: $N_{pi} \sim \Gamma(a_i, \theta_i)$, where the distribution is described by two gene-specific statistical parameters: the scale parameter θ_i and the shape parameter a_i [13, 18, 19]. In what follows, we will drop the explicit gene subscript i for readability. These two statistical parameters have clear biological interpretations in terms of the *Telegraph Model*, the kinetic model for the central dogma [19]. The scale parameter is proportional to the *translation efficiency* (ε), the mean number of proteins translated from each message transcribed for gene i: $\theta = \varepsilon \ln 2$, and the shape parameter (a) is proportional to the message number: $a = \mu_m / \ln 2$, where the *message number* μ_m is defined as the mean number of messages transcribed per cell cycle for gene i. See the Supplementary Material Sec. A 1 a for a detailed description of the model and the relationships between the model parameters.

In terms of these gene-specific central dogma parameters, the *protein number* (μ_p), defined as the mean protein number per cell at cell birth, and the coefficient of variation (CV_p^2) are:

$$\mu_p = \mu_m \varepsilon, \tag{1}$$

$$\mathrm{CV}_p^2 = \frac{\ln 2}{\mu_m}, \tag{2}$$

for gene i. The protein number can be understood as the product of the gains of a two-stage amplification process: transcription followed by translation. (See Fig. 2A.) However, under typical biological conditions, the noise is dominated by transcription and therefore is inversely proportional to the message number μ_m [19].

Next, we consider the metabolic load on the cell as a consequence of protein expression. In the absence of cell cycle arrest, we will assume that the cell cycle duration τ is proportional to the overall metabolic load of the messages, proteins, and other cellular components [20]. Focusing on the metabolic load of a particular gene, the inverse relative growth rate is:

$$\frac{k_0}{k} = 1 + \frac{\lambda + \varepsilon}{N_0} \mu_m, \tag{3}$$

where k_0 is the growth rate in the absence of the metabolic load of gene i, N_0 is the total metabolic load of all genes (in protein equivalents) and λ is the metabolic message cost. (See Supplementary Material Sec. A 2 for a detailed derivation.) The λ-term represents the metabolic cost of transcription and the ε-term represents the metabolic cost of translation of gene i. Although the global parameters N_0 and λ provide an intuitive representation of the model, the relative growth rate depends on fewer parameters. We define the relative load as $\Lambda \equiv \lambda / N_0$ as the ratio of the metabolic load of a single message to the total load and the relative translation efficiency $E \equiv \varepsilon / N_0$ as the ratio of the number of proteins translated per message to the total metabolic load N_0. In *E. coli*, we estimate that both Λ and E are roughly 10^{-5} and they are smaller still for eukaryotic cells. (See Supplementary Material Sec. A 11.)

The final task is to link protein expression with cellular function. Motivated by the concept of rate-limiting reactants [21] as well as single-cell growth rate measurements [22], we will propose that the essential processes of the cell have a threshold-like dependence on each essential protein: We will assume that each essential protein i has a critical gene-specific threshold number n_p such that growth arrests below this critical number. (See Fig. 2CDE.) In our analysis, we will treat the thresholds n_p as gene-specific unknown parameters that could be determined experimentally. It will usually be more convenient to work in terms of the threshold fraction, $\phi_p \equiv n_p / N_0$, which can be interpreted as the threshold protein fraction required for growth.

We have previously analyzed a model in which the duration of cell-cycle phases (or periods) are of stochastic duration [23]. In particular, we have already considered the case of stochastic cell-cycle arrest, which can

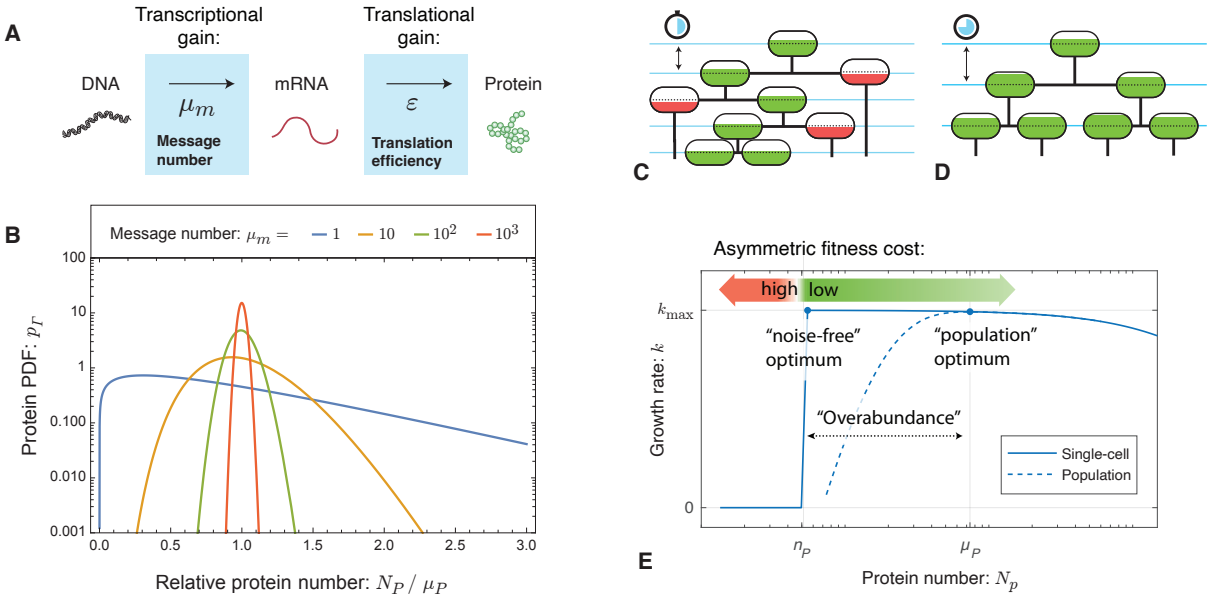

FIG. 2. **Panel A: Two-stage amplifier model of expression.** The central dogma describes a two-step stochastic process by which genes are first transcribed and then translated. The transcription process generates an average of μ_m, the message number, of messages per cell cycle. Translation generates an average of ε, the translation efficiency, proteins per message transcribed. **Panel B: Gene expression noise.** The protein number is observed to be Gamma-distributed due to the stochasticity of the central dogma processes. The noise is determined by the message number μ_m. For highly-transcribed genes (*e.g.* $\mu_m = 10^3$), the distribution is tightly distributed about its mean; however, for low-expression genes (*e.g.* $\mu_m = 1$), the distribution is extremely wide, resulting in many cells with extremely low protein abundance. **Panel C & D: The RLTO Model.** A schematic cell lineage tree is shown during exponential growth. The cell fill represents an essential protein expression level. The dotted line represents the threshold number n_p required for cell growth. **Panel C:** Reducing the expression level reduces the metabolic load (the spacing between blue lines); however, below-threshold cells arrest (red). **Panel D:** Increasing protein expression increases the metabolic load (the spacing between blue lines); however, all cells are above threshold. **Panel E: Protein overabundance optimizes growth rate.** The solid blue line represents the growth rate as a function of protein number for a single cell. Below the threshold level n_p, there is no growth. Above the threshold, the growth rate decreases slowly as a consequence of the metabolic load. The dashed blue line represents the population growth rate as a function of mean protein level. The growth rate is optimized at μ_p, far above the threshold due to the cell-to-cell variation in protein number. The asymmetric fitness landscape causes the optimal protein expression level to be overabundant.

Model parameter name:	**Symbol:**	**Units:**	**Global/ Specific:**	**Description:**
Total metabolic load	N_0	Protein molecules	global	Total metabolic load in protein molecule equivalents
Message cost	λ	Protein molecules	global	Metabolic cost per message transcribed per cell cycle
Relative load	$\Lambda \equiv \lambda/N_0$	Number	global	Message cost relative to total load
Threshold number	n_p	Protein molecules	gene i	Protein number required for function
Threshold fraction	$\phi_p \equiv n_p/N_0$	Number	gene i	Protein fraction required for function
Message number	μ_m	Number	gene i	Total number of messages transcribed per cell cycle
Translation efficiency	ε	Protein molecules	gene i	Number of protein made per message
Relative translation efficiency	$E \equiv \varepsilon/N_0$	Number	gene i	Number of protein made per message relative to total load
Protein number	$\mu_p \equiv \mu_m\varepsilon$	Protein molecules	gene i	Protein number
Protein fraction	$\Phi_p \equiv \mu_p/N_0$	Number	gene i	Protein number relative to total load
Overabundance	$o \equiv \mu_p/n_p$	Number	gene i	Protein number relative to threshold number

TABLE I. **Summary parameters.** A summary of model parameters is given for the RLTO model. The top three parameters are global. The middle and bottom sets of parameters are all gene specific (Gene i). The top and middle parameter sets are assumed to be fixed in the model; whereas the bottom set are optimized to maximize growth rate. Per gene i, there are two independent parameters (μ_m and ε) and one fixed, gene-specific parameter (n_p). Model parameters appearing with a hat are optimal values (*e.g.* optimal message number: $\hat{\mu}_m$).

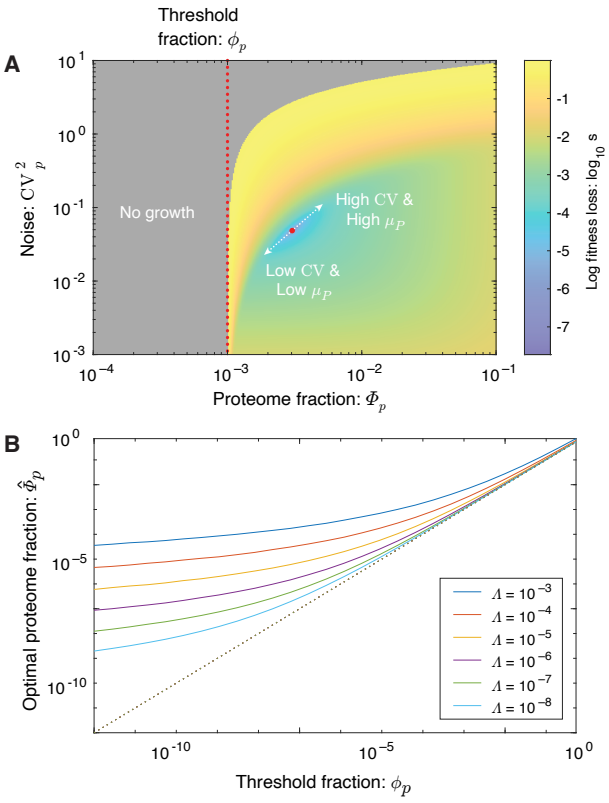

FIG. 3. **Panel A: Fitness landscape.** The fitness loss ($s \equiv \ln k_{\max}/k$) is shown as a function of noise (CV_p^2) and protein fraction. The red dotted curve shows the threshold ϕ_p, the red dot represents the optimum fitness, and the distance between these levels represents the overabundance ($o = \Phi_p/\phi_p$). Robustness to noise drives the optimum expression level significantly above the threshold ϕ_p. The low and higher noise strategies discussed above are shown with the white-dotted line. **Panel B: Optimal protein fraction.** The optimal protein fraction is shown as a function of the threshold fraction. To prevent growth arrest, the protein fraction is always larger than the threshold fraction, leading to overabundance. However, the overabundance grows rapidly as the threshold is reduced. The optimal protein fraction also depends on the relative load Λ. Higher relative load leads to a smaller protein fraction at a fixed threshold.

be computed analytically. Assuming the expression of each protein is independent, the protein number in subsequent cell cycles are uncorrelated [13] and the failure probability is small, the growth rate of the RLTO model can also be computed analytically. (See Supplementary Material Sec. A 2 a.) The relative growth rate is:

$$\ln \frac{k}{k_0} = -(\Lambda + E)\mu_m - \frac{1}{\ln 2}\gamma\left(\frac{\mu_m}{\ln 2}, \frac{\phi_p}{E \ln 2}\right), \tag{4}$$

where γ is the regularized incomplete gamma function, which is the CDF of the gamma distribution and represents the probability of arrest due to gene i. Eq. 4 represent an explicit analytic model for cell fitness that accounts for growth robustness to noise, metabolic load, and high multiplicity. In the RLTO model, the relative growth rate depends only on a single global parameter: the relative metabolic load Λ and three gene-specific parameters: the threshold fraction ϕ_p, the message number μ_m and the relative translation efficiency E. We propose that the cell is regulated to optimize the message number and translation efficiency to maximize the growth rate. The model parameters are summarized in Tab. I.

The fitness landscape of a trade-off. The fitness landscape predicted by the RLTO model for representative parameters is shown in Fig. 3. (A detailed description of the model parameters for all figures is provided in the Supplementary Material in Tab. III.) The figure displays a number of important model phenomena: There is no growth for a protein fraction Φ_p, defined as $\Phi_p \equiv \mu_p/N_0$, below the threshold value ϕ_p, and for high noise, Φ_p must be in significant excess of ϕ_p. Rapid growth can be achieved by the two mechanisms described in Fig. 1: (i) high expression levels (Φ_p) are required for high noise amplitude (CV_p^2) or (ii) lower expression levels coupled with lower noise. This trade-off leads to a ridge-like feature of nearly optimal models represented by the dotted white line. The optimal fitness corresponds to a balance between increasing the mean protein expression (Φ_p) and decreasing noise (CV_p^2). This optimal central dogma program strategy leads to significant overabundance. (See Fig. 3B.)

RLTO predicts protein overabundance. How does this optimal strategy depend on the threshold fraction (ϕ_p) and the relative load (Λ)? To understand the phenomenology of the model, we optimize the growth analytically. Although the threshold fraction ϕ_p is intuitive from a mechanistic perspective, it is less convenient from an experimental perspective. We therefore define the overabundance o, the protein number relative to the threshold:

$$o \equiv \Phi_p/\phi_p, \tag{5}$$

where Φ_p is the protein fraction and ϕ_p is the threshold fraction required for growth (*e.g.* [16]).

To maximize the growth rate, we set the partial derivatives of Eq. 4 with respect to message number and translation efficiency to zero and then solve the coupled transcendental equations for the optimal message number $\hat{\mu}_m$ and relative translation efficiency $\hat{E}$. As shown in Supplementary Material Sec. A 3, the equation for the optimal message number is:

$$\Lambda \ln 2 = -\frac{\partial}{\partial \hat{\mu}_m}\gamma\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o} \ln 2}\right). \tag{6}$$

The resulting equation depends on only a single parameter: the relative load Λ, and is independent of the optimal relative translation efficiency.

The predicted relation between the optimal message number ($\hat{\mu}_m$) and overabundance ($\hat{o}$) is shown in Fig. 4A. The RLTO model generically predicts that the optimal protein fraction is overabundant ($o > 1$). For

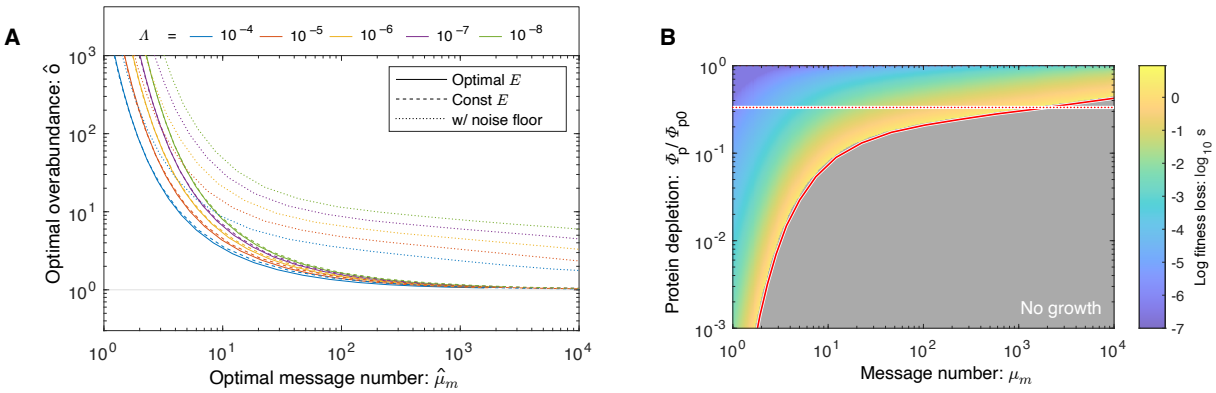

FIG. 4. **Panel A: Overabundance as a function of message number.** In the limit of high message number (μ_m), the noise is small, and the optimal expression level is close to the threshold ϕ_p, leading to minimal overabundance ($o \to 1$); however, as the message number approaches unity, the noise is comparable to the mean, driving vast overabundance ($o \gg 1$). Overabundance increases as the relative metabolic load decreases, and is essentially identical for constant (dashed) versus optimized translation efficiency (solid). To model overabundance in bacterial cells, we included the contribution of the noise floor with fixed translation efficiency (dotted curves), which increases overabundance, especially for high expression genes. **Panel B: Optimal expression levels are buffered.** The predicted fitness loss as a function of protein depletion level and message number for bacterial cells (including the noise floor). Due to the overabundance phenomenon, all proteins are buffered against depletion, but low-expression genes are particularly robust due to higher overabundance. The solid red line represents $1/o$, and predicts the range of depletion values for which cell growth is predicted. The dotted red line represents a three-fold depletion.

highly-transcribed genes ($\mu_m \gg 1$) like ribosomal genes, the overabundance is predicted to be quite small ($o \approx 1$); however, for message numbers approaching unity, the overabundance is predicted to be extremely high ($o \gg 1$). At a quantitative level, the relation between optimal overabundance and message number does depend on the relative load (Λ), but its phenomenology is qualitatively unchanged over orders of magnitude variation in Λ.

RLTO predicts larger overabundance in bacteria. There are two distinctive features of bacterial cells that could affect the model predictions: (i) the translation efficiency is constant [24] and bacterial gene expression is subject to a large-magnitude noise floor that increases the noise for high-expression genes [13]. The optimization of message number at fixed translation efficiency does result in a slightly modified optimization condition for the message number (Supplementary Material Sec. A 4 a); however, the predicted overabundance is only subtly perturbed (Fig. 4A). In contrast, the noise floor increases the predicted overabundance, especially for high-expression proteins. As a result, the RLTO model predicts that the vast majority of bacterial proteins are expressed in significant overabundance. (See Fig. 4A.)

Overabundance is a robust prediction. Is the predicted overabundance an artifact of the RLTO model? We hypothesize that overabundance might be a consequence of growth arrests for protein numbers below the threshold. To test whether growth arrest is required for the observed RLTO model phenomenology, we consider a *slow-growth model* where the growth rate smoothly decreases to zero at low protein number, but does not arrest. (See the Supplementary Material Sec. B for a detailed description of the models.) This numerical analysis reveals that the slow-growth and RLTO models had qualitatively identical phenomenology (*i.e.* $o \gg 1$). We therefore concluded that explicit arrest is not required to make overabundance optimal.

Next, we hypothesized that overabundance is the consequence of the highly-asymmetric fitness landscape (Fig. 2E). In both the slow-growth and RLTO models, the growth rate rapidly decreases for underabundance but decays gradually for overabundance. To explore the consequences of asymmetry, we analyzed a *symmetric model* with a symmetric fitness landscape. (See the Supplementary Material Sec. B for a detailed description of the models.) Consistent with the hypothesis that growth rate asymmetry is the key mathematical mechanism to drive overabundance, we observe that the symmetric model was optimized very close to the noise-free optimum protein number (*i.e.* the model did not predict overabundance and $o = 1$). We conclude that fitness asymmetry, but not growth arrest, is the characteristic that drives overabundance in the models.

RLTO predicts proteins are buffered to depletion. A principle motivation for our analysis is the observation that many protein levels appear to be buffered. To explore the prediction of the RLTO model for protein depletion, we first computed the optimal message numbers and translation efficiencies for a range of protein thresholds. To model the effect of protein depletion, we computed the change in growth rate as function of pro-

tein depletion (equivalent to a reduction of the translation efficiency relative to the optimum.) The growth rate is shown in Fig. 4B for the RLTO model with parameters representative of a bacterial cell. (See Supplementary Material Sec. A 4 b.)

In general, the RLTO model predicts that protein numbers have very significant robustness (*i.e.* buffering) to protein depletion. This is especially true for low expression proteins that are predicted to have the largest overabundance. For these genes, even a ten-fold depletion leads to very subtle reductions in the growth rate. For a three-fold reduction in the growth rate, only the very highest-expression genes (*e.g.* ribosomal genes) are expected to lead to qualitative phenotypes.

Overabundance is observed in a range of experiments. The RLTO Model predicts that all essential proteins are overabundant. Although this result is potentially surprising, it is in fact consistent with many studies. For instance, Belliveau *et al.* have recently analyzed the abundance of a wide range of metabolic and other essential biological processes, and conclude that protein abundance appears to be in significant excess of what is required for function [9]. Likewise, CRISPRi approaches have facilitated the characterization of essential protein depletion. The qualitative results from these experiments are consistent with overabundance: Large-magnitude protein depletion is typically required to generate strong phenotypes [15, 17, 25]. In particular, Peters *et al.* engineered a complete collection of CRISPRi essential-gene depletion constructs in *Bacillus subtilis*. Importantly, when *dcas9* is constitutively expressed, these constructs deplete essential proteins about three-fold below their endogenous expression levels [15]; however, roughly 80% grew without measurable fitness loss in log-phase growth despite the depletion. When grouped by functional category, only ribosomal proteins were found to have statistically significant reductions in fitness [15]. As shown in Fig. 4B, the RLTO model predicts that all but the highest expression proteins are expected to show minimal fitness reductions in response to a three-fold depletion of essential enzymes.

Although this qualitative picture of essential protein overabundance is clear, there has yet to be a quantitative and detailed measurement of protein overabundance, and in particular, an analysis of the relationship between protein overabundance and message number. We will present an explicit experimental test of this prediction elsewhere [22].

RLTO predicts a one-message transcription threshold. The RLTO model predicts protein overabundance, but is there a clear transcriptional signature? To analyze this question, we define the message threshold $n_m \equiv \mu_m/o$. (This parameterization is convenient since it is independent of the translation efficiency.) We can then analyze the relation between optimal message number and threshold message number, as shown in Fig. 5A. The model predicts that even for genes that have extremely

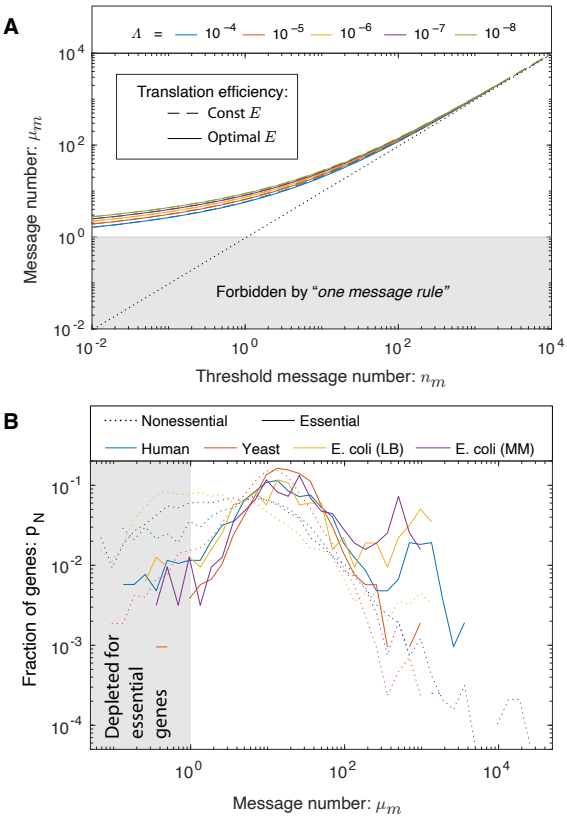

FIG. 5. **Panel A: One-message-rule for essential genes.** For highly transcription genes (high μ_m), little compensation for noise is required and the optimal message number tracks with the threshold message number n_m. However, as the threshold message number approaches one ($n_m \to 1$), the noise is comparable to the mean, and the optimal message number μ_m increases to compensate for the noise. As a result, a lower threshold of roughly one message per cell-cycle is required for essential genes. This threshold is predicted for both fixed (dashed) and optimized translation efficiency (solid). The threshold is weakly dependent on relative load Λ. **Panel B: A one message threshold is observed in three evolutionarily-divergent organisms.** As predicted by the RLTO model, essential, but not nonessential genes, are observed to be expressed above a one message per cell-cycle threshold. All organisms have roughly similar distributions of message number for essential genes, which are not observed for message numbers below a couple per cell cycle.

small threshold message numbers (*e.g.* $n_m = 10^{-2}$), the optimal message number stays above one message transcribed per cell cycle. Qualitatively, expressing messages below this level is simply too noisy even for proteins needed at the lowest expression levels. The model therefore predicts a lower floor on transcription for essential genes of one message per cell cycle.

A lower threshold is observed for message number. To identify a putative transcriptional floor, we consider the central dogma in three model organisms: the bacterium *E. coli*, *Saccharomyces cerevisiae* (yeast) and *Homo*

sapiens (human). For *E. coli*, we consider both rapid and slow growth conditions. We analyze three different transcriptional statistics for each gene: transcription rate (β_m), cellular message number ($\mu_{m/c}$), defined as the average number of messages instantaneously, and message number (μ_m), defined as the number of messages transcribed in a cell cycle. Analysis of these organisms explores orders-of-magnitude differences in characteristics of the central dogma, including total message number, protein number, doubling time, message lifetime, and number of essential genes. (See Supplementary Tab. IV.)

In the previous section, we hypothesized that cells must express essential genes above some threshold message number for robust growth; however, we expect to see that non-essential genes can be expressed at much lower levels since growth is not strictly dependent on their expression. A detailed description of the estimation of transcription rates, cellular message number, and message number, and a detailed description of the sources of the data are provided in the Supplementary Material Sec. C. To identify the putative transcriptional threshold, we generated histograms of each statistic in each organism and growth condition.

As expected, there does not appear to be any consistent lower threshold between *E. coli*, yeast, or human transcription, either as characterized by the transcription rate (β_m) or the cellular message number ($\mu_{m/c}$). (See Supplementary Fig. 17.) However, as predicted by the RLTO model, there is a consistent lower limit on message number (μ_m) of roughly one message per cell cycle for essential genes. (See Fig. 5B.) This floor is consistent not only between *E. coli*, growing under two different conditions, but also between the three highly-divergent organisms: *E. coli*, yeast and human. We will conservatively define the minimum message number as:

$$\mu_m \geq 1, \tag{7}$$

and summarize this observation as the *one-message-rule* for essential gene expression.

There are a small number of essential genes that appear to be expressed below this threshold. For instance, in *E. coli* there are 10 genes. When we investigated these genes in more detail, only one was consistently classified *essential* between multiple studies, and this last gene was a prophage gene. (See Supplementary Material Sec. A 6 for the gene list, the analysis, and a more detailed discussion.)

In addition to the common floor for essential genes, there is a commonality in message number distribution between organisms. To appreciate the significance of this observation, we note that no such similarity is seen in the distribution of transcription rates or cellular message numbers. (See Supplementary Fig. 17.) This similarity emphasizes both the conservation of the central dogma regulatory program and the importance of message number as the key transcriptional statistic.

In contrast to essential genes, non-essential genes can be expressed with message numbers below the threshold. (See Fig. 5B.) However, it is important to emphasize that the difference between essential and non-essential genes is less significant than it initially appears. These non-essential genes include those that are inducibly expressed but are not induced under the growth conditions studied (*e.g.* the *lac* operon in *E. coli*). The similarity between essential and non-essential gene expression can be seen in the yeast message number distributions: The essential and non-essential curves are very similar. We believe that the principal analytical significance of the analysis of essential genes is to confine our analysis to genes that are transcriptionally active under the conditions studied.

Prediction of the optimal load ratio. The two-stage amplification of the central dogma implies that the expression and noise levels can be controlled independently by the balance of transcription to translation. How does the cell achieve high and low gene expression optimally, and how does this strategy depend on the message cost?

To understand the optimization, we first define the load ratio R for a gene as the metabolic cost of translation relative to transcription:

$$R \equiv \frac{\mu_p}{\lambda \mu_m} = \frac{\varepsilon}{\lambda}. \tag{8}$$

In the Supplementary Material Sec. A 3, we show that the optimal load ratio is:

$$\hat{R} \equiv \frac{1}{\Lambda \ln 2}\, p_\Gamma(\hat{\mu}_m | \tfrac{\hat{\mu}_m}{\ln 2}, \tfrac{1}{\delta \ln 2}), \tag{9}$$

where p_Γ is the PDF of the gamma distribution. The optimal load ratio is shown in Fig. 7A.

The dependence of the optimal load ratio $\hat{R}$ on Λ is extremely weak, but it is strongly dependent on message number. As a result, for low transcription genes ($\mu_m \ll 10$), the metabolic load is predicted to be dominated by transcription; whereas, for highly transcribed genes ($\mu_m \gg 10$), the metabolic load is dominated by translation. These predictions are robust since they are independent of the relative load Λ.

Translation efficiency is predicted to increase with transcription. Now that we have defined the optimal load ratio, the equation for optimal translation efficiency can be written concisely:

$$\hat{\varepsilon} = \lambda \hat{R} \quad \text{or} \quad \hat{E} = \Lambda \hat{R}, \tag{10}$$

where $\hat{R}$ depends weakly on the relative load Λ. The RLTO model predicts that optimal partitioning of amplification between transcription (gain μ_m) and translation (gain ε) has two important qualitative features: (i) As the message cost (λ) rises, the optimal translation efficiency increases in proportion. (ii) The optimal translation efficiency is also approximately proportional to message number ($\hat{\varepsilon} \propto \mu_m$). (See Fig. 7A.) Therefore, the RLTO model predicts that low expression levels should be achieved with low levels of transcription and translation, whereas high expression genes are achieved with

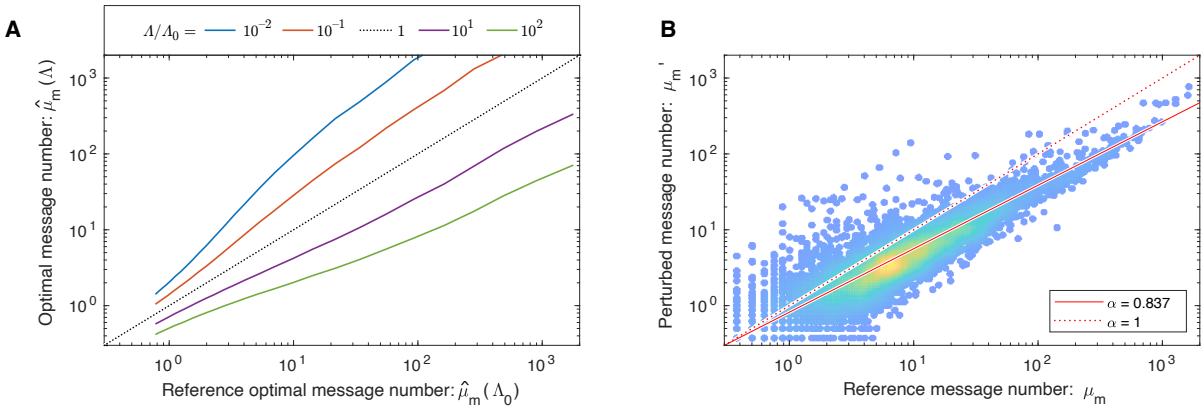

FIG. 6. **Panel A: Message number decreases with increased relative load Λ.** The optimal message number responds to changes in the message cost. The RLTO model predicts an approximate power-law relation (linear on a log-log plot) between message numbers. **Panel B: A power-law relation is observed.** To test whether central dogma regulation would adapt dynamically as predicted, we analyzed the relation between the yeast transcriptome under reference conditions and phosphate depletion (perturbed), which increases the message cost [26]. (Data from Ref. [27].) As predicted by the RLTO model, a global change in regulation is observed, which generates a power-law relation with scaling exponent $\alpha = 0.837 \pm 0.01$. The observed exponent is smaller than one, as predicted by an increased relative load Λ.

high levels of both. We call this relation between optimal transcription and translation *load balancing*.

Message number also responds to message cost. We will first focus on analyzing the implications of the message cost dependence in Eq. 10. At a fixed load ratio, Eq. 10 clearly implies that the translation efficiency increases as the message cost λ increases; however, the message number (and load ratio) also respond to compensate to changes in λ. To probe the dependence on message cost in an experimentally relevant context, consider optimal message numbers in a reference condition (relative load Λ_0) relative to a second perturbed condition (relative load Λ). The predicted relation between the optimal messages numbers is shown in Fig. 6A. The resulting relation between the optimal message numbers is roughly linear on a log-log plot, predicting the approximate power-law relation:

$$\hat{\mu}_m(\Lambda) \propto \hat{\mu}_m(\Lambda_0)^{\alpha}, \tag{11}$$

describing a non-trivial global change in the regulatory program.

RLTO predicts the yeast global regulatory response. To test the RLTO predictions, we compared the relative message numbers for yeast growing under phosphate depletion, which increases the message cost [26], to a reference condition [27]. As predicted, the relative transcriptome data was well described by a power law (Eq. 11) and the observed slope was smaller than one: $\hat{\alpha} = 0.837 \pm 0.001$, as predicted by the increased message cost. See Fig. 6B.

The observation of this large-scale regulatory change has an important implication: This response supports a nontrivial hypothesis that the RLTO model not only can predict how the cell is optimized in an evolutionary sense, but can predict global regulatory responses as well.

Parameter-free prediction of proteome fraction. We now turn our focus to an analysis of the implications of the message number dependence of the translation efficiency (Eq. 10). The most direct test of this prediction is measuring the relation between proteome fraction and message number. The RLTO model predicts protein fraction:

$$\hat{\Phi}_p = \hat{E}\hat{\mu}_m \propto \hat{\mu}_m^2, \tag{12}$$

where μ_m is the observed message number and the optimal relative translation efficiency is predicted by Eq. 10. The proportionality is only approximate but gives important intuition for how protein number depends on message number in the RLTO model, in contrast to a constant-translation-efficiency model: $\Phi_p \propto \mu_m$. To compare these predictions to protein abundance measurements, we will renormalize the protein fraction to be defined relative to total protein number rather than N_0. This renormalization eliminates the Λ dependence to result in a parameter-free prediction of the proteome fraction.

RLTO predicts proteome fractions in eukaryotic cells. To test the RLTO predictions, we compare observed proteome measurements in three evolutionarily divergent species, *E. coli* [24], yeast [28] and mammalian cells [29], to two models: the RLTO and the constant-translation-efficiency models. The results of the parameter-free predictions are shown in Fig. 7BCD. The RLTO model clearly captures the global trend in the proteome-fraction message-number relation in eu-

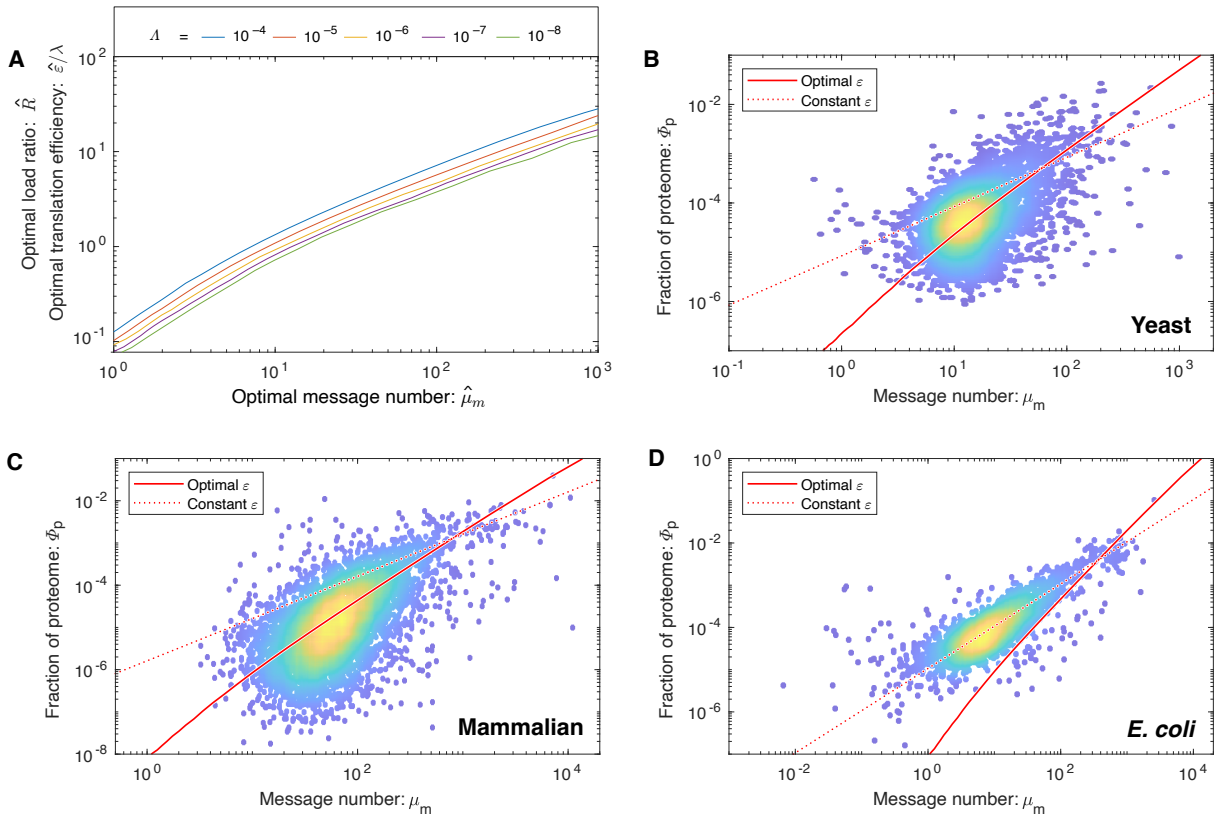

FIG. 7. **Panel A: Load balancing: the optimal translation efficiency increases with message number and cost.** The ratio of the optimal translation efficiency ($\hat{\varepsilon}$) to the message cost (λ) is roughly independent of the relative load (Λ); therefore, the optimal translation efficiency is expected to be proportional to the message cost. The translation efficiency is also roughly proportional to the optimal message number $\hat{\mu}_m$. The ratio ε/λ has a second interpretation: the load ratio R. R is defined as the metabolic cost of translation over transcription of the gene. **Gene proteome fraction versus message number for three model systems.** Two models for relation between gene proteome fraction and message number are compared with observations: The RLTO model makes a parameter-free prediction of the optimal relation (*optimal*, solid red line). A second competing model is constant translation efficiency (*constant*, dotted red line). **Panel B: Yeast proteome fraction.** The RLTO prediction (solid) is superior to the constant-translation-efficiency prediction (dashed). **Panel C: Mammalian proteome fraction.** The RLTO prediction (solid) is superior to the constant-translation-efficiency prediction (dashed). **Panel D: *E. coli* proteome fraction.** In contrast, the constant-translation-efficiency prediction (dashed) is superior to RLTO prediction (solid).

karyotic cells, but not in *E. coli*, where the constant-translation-efficiency models better describes the data.

What gives rise to the spread of the data around the optimal protein fraction, and why does the RLTO model fail to describe *E. coli*? In the Supplementary Material, we analyze a number of refinements to the RLTO model. Motivated by the data spread, we investigate models in which there are gene-specific supplemental loads associated with proteins (*e.g.* toxicity) and models in which gene and protein length are treated explicitly (Supplementary Material Sec. A 8 a). We conclude that the observed data could be explained by supplemental load, but not gene and protein length. Motivated by the failure of the optimal translation efficiency to describe the proteome fraction in *E. coli*, we consider two different mechanisms which limit translation (Supplementary Material Sec. A 9). The *E. coli* data is consistent with a ribosome-per-message limit, as proposed by Hausser *et al.* [14].

RLTO model predicts non-canonical noise scaling. The predicted scaling of the optimal translation efficiency with message number has many important implications, including on the global characteristics of noise. Based both on theoretical and experimental evidence, it is widely claimed that gene-expression noise should be inversely proportional to protein abundance [13, 30]:

$$\mathrm{CV}_p^2 \propto \mu_p^{-1}, \tag{13}$$

for low-expression proteins, as observed in *E. coli* [13]; however, the more fundamental prediction is that the noise is inversely proportional to the message number (Eq. 2). In *E. coli*, the translation efficiency is roughly constant (*i.e.* $\hat{\mu}_p \propto \hat{\mu}_m$) and therefore Eq. 2 is consistent with the canonical noise model (Eq. 13). However, in eukaryotes, the translation efficiency grows with mes-

sage number (*i.e.* $\hat{\mu}_p \propto \hat{\mu}_m^2$). If we substitute this proportionality into Eq. 2, we predict the non-canonical noise scaling:

$$\mathrm{CV}_p^2 \propto \mu_p^{-1/2}, \tag{14}$$

for eukaryotic cells.

Non-canonical noise scaling is observed in yeast. To test the RLTO model predictions for noise scaling, we reanalyze the dataset collected by Newman *et al.*, who performed a single-cell proteomic analysis of yeast by measuring the abundance of fluorescent fusions by flow cytometry [12]. Since the competing models (Eqs. 13 and 14) make different scaling predictions, we first apply a statistical test to determine whether the observed scaling is consistent with the canonical model (Eq. 13). We consider the null hypothesis of canonical model (Eq. 13) and the alternative hypothesis with an unknown scaling exponent. To test the models, we performe a null hypothesis test. (A detailed description of the statistical analysis, which include the contribution of the noise floor, is given in the Supplementary Material Sec. D.) We reject the null hypothesis with a p-value of $p = 6 \times 10^{-36}$. The observed scaling exponent is $\hat{a} = -0.57 \pm 0.02$, which is close to our predicted estimated exponent from the RLTO model ($-\frac{1}{2}$).

Prediction of noise from protein-message relation. By combining the noise model (Eq. 2) with a protein-message abundance relation, the relation between protein abundance and noise can be predicted without additional fitting parameters. To test this prediction, we will compare three competing models: (i) the RLTO model, (ii) an empirical protein-message abundance model, and (iii) the constant-translation-efficiency model. (See the Supplementary Material Sec. D for a detailed description of the analysis.) The fit of the competing protein-message abundance models are shown in Fig. 8A. Using each model, we can now predict the relation between protein abundance and noise without additional fitting parameters. The predictions of the three competing models are compared to the experimental data in Fig. 8B.

In both its ability to capture the protein abundance and predict the noise, the RLTO model vastly outperforms the constant-translation-efficiency model. The purely empirical model that best captures the protein abundance data, due to directly fitting both the y-offset and slope, also performs best in predicting the noise. It is important to emphasize that the prediction of the noise in all models is non-trivial since there are no free parameters fit, once the protein abundance relation is determined. We therefore conclude that the noise model (Eq. 2) quantitatively predicts the observed noise from the message number and that eukaryotic noise has non-canonical scaling due to load balancing.

Load balancing reduces the noise of the noisiest genes. Now that we understand the consequences of load balancing on the scaling of the noise, we return to comment on its significance. Why is load balancing predicted to be optimal by the RLTO model? Fig. 8B qualitatively illustrates its rationale: Load balancing (green and purple curves) decreases the noise relative to the constant-translation-efficiency model (orange curve) for

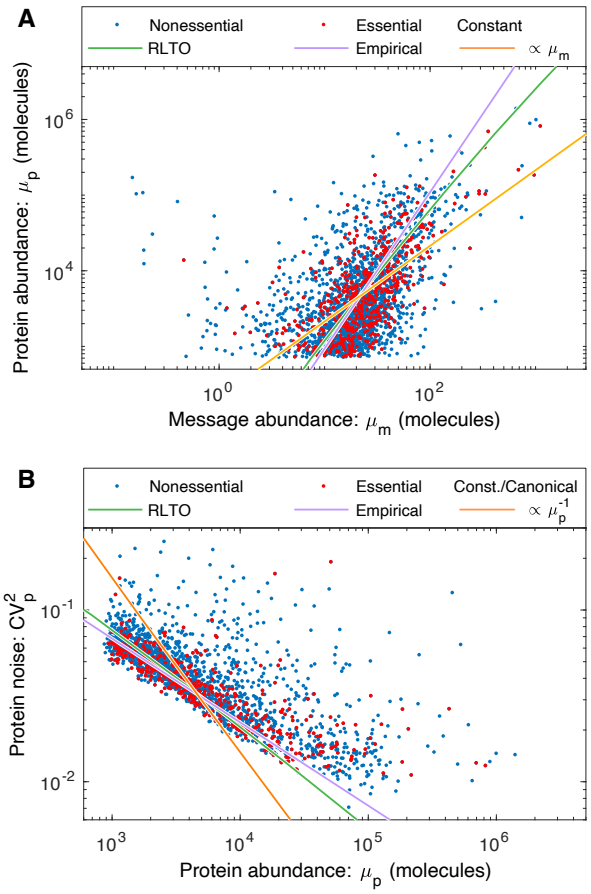

FIG. 8. **Panel A: Three competing models for protein abundance in yeast.** The empirical model (purple) fits the slope and the y offset. The RLTO (green) and constant-translation-efficiency (orange) models fit a parameter corresponding to the y offset only. As discussed in the analysis of the proteome fraction, the RLTO model qualitatively captures the scaling of the protein abundance with message number better than the constant translation efficiency model; however, the predicted fit does not correspond to the optimal power law, which is represented by the empirical model. The protein abundance has a cutoff near 10^1 due to autofluorescence [12]. **Panel B: Predictions of the noise-protein abundance relation.** Using each competing protein abundance model, the noise-protein abundance relation can be predicted using Eq. 2. The canonical noise model (Eq. 13) fails to capture even the scaling of the noise. In contrast, both the RLTO and empirical models quantitatively predict both the scaling and magnitude of the noise. The empirical model has the highest performance, presumably due to its two-parameter fit to the protein abundance in Panel A. A fit accounting for the noise floor is shown in Supplementary Material Fig. 19.

low expression genes. These low-expression genes are the noisiest. Load balancing therefore reduces the noise of the noisiest genes

DISCUSSION

To analyze the consequences of the trade-off between metabolic load and growth robustness, we developed a novel, analytically-tractable approach: the Robustness-Load Trade-Off (RLTO) model. The model includes an important theoretical innovation: It is tractable beyond the small-fitness-perturbation limit implemented in previous analyses (*e.g.* [14]). As a result, we can explore the consequences of a highly-asymmetric fitness landscape for the first time. This new model describes a rich phenomenology that is absent from previous models.

Understanding the rationale for overabundance. Essential protein overabundance is the signature prediction of the RLTO model. Its mathematical rationale is the highly-asymmetric fitness landscape. To understand why we expect this rationale to be generic, consider the form of the optimization condition for message number (Eq. 6). The growth rate is maximized when the probability of slow-growth (*e.g.* arrest) is roughly equal to the relative load of adding one more message. Since the cell makes roughly 10^5 messages per cell cycle, the relative load is extremely small and therefore the probability of slow growth must be as well. Making this probability very small requires vast overabundance for the inherently-noisy, low-expression proteins. Should this strategy surprise us? No. We routinely use this strategy in our own lives. For instance, when baking an apple pie, we keep cheap ingredients, needed in small quantities, in vast overabundance (*e.g.* salt); whereas, the more expensive ingredients, needed in higher quantities (*e.g.* apples), are bought only as needed. (See Fig. 9A.)

Implications of overabundance for inhibitors. The generic nature of overabundance, especially for low-expression proteins, has important potential implications for the targeting of these proteins with small-molecule inhibitors (*e.g.* drugs). For highest expression proteins, like the constituents of the ribosome, relatively small decreases in the active fraction (*e.g.* a three-fold reduction) are expected to lead to growth arrest [15]. This may help explain why inhibitors targeting translation make such effective antimicrobial drugs. (See Fig. 4B.) However, we predict that the lowest expression proteins require a much higher fraction of the protein to be inactivated, with the lowest-expression proteins expected to need more than a 100-fold depletion. This predicted robustness makes these proteins much less attractive drug targets [31].

The role of regulatory feedback. Can transcriptional-regulatory feedback circumvent the need for overabundance by eliminating noise? Transcriptional regulation does not appear to eliminate protein fluctuations [18]. For instance, the distribution of protein number has been characterized for about quarter of all genes in *E. coli* [13] and the observed distribution in protein numbers are very well described by the Gamma distribution, including essential gene products [13]. Similarly, a proteome-scale analysis of noise in yeast also seems consistent with the lower bound on noise predicted by the telegraph model. (See Fig. 8.) No putative set of low-noise essential genes appears to exist, although essential genes do appear to reduce noise by two mechanisms: obeying the *one-message-rule* in all organisms and by load balancing in eukaryotic cells.

Implications for non-essential genes. In our analysis, we have focused on essential genes in order to motivate the growth-threshold in the RLTO model. To what extent do non-essential genes share the same optimization? In support of the proposal that RLTO optima describe non-essential genes is the success of the model in predicting the translation efficiency for all genes, not just essential genes. (See Fig. 7.) Furthermore, the definition of a gene as *essential* depends on context: For instance, in the context of *E. coli* growth on lactose, the gene *lacZ* is essential, although it is non-essential on other carbon sources [32]. Under growth conditions where the *lacZ* gene is essential, we predict that LacZ should be overabundant. The interesting aspect to *lac* operon expression is that it is not constitutive, but induced in the presence of lactose. It is therefore natural to predict that the *lac* operon is either *off* or *on-and-overabundant*. This biphasic behavior has long been known: Novick and Weiner reported that *enzyme induction [is] an all-or-none phenomenon* [33]. Furthermore, the *lac* operon expression does appear to be overabundant when expressed, since a *metabolic memory effect* is observed, in which multiple generations of cell-growth-induced protein dilution are required before cells lose their adaptation for growth on lactose [10]. In analogy to the *lac* operon, we expect all gene products, most especially those with low expression, to be overabundant, under conditions where their activity is essential.

Environmental fluctuations. Environmental fluctuations have been proposed as the principal rationale for overabundance [3]. In short, cells express protein to bet-hedge against starvation [3] or changes in the carbon source, *etc.* [10]. There are some similarities between these environmental fluctuation models and the RLTO model: In both models, it is a fluctuation-based mechanism that drives overabundance; however, there are important distinctions. In the environmental fluctuation model, there is a trade-off between log-phase fitness and the rapidity of adaptation [3]; whereas in the RLTO model, overabundance corresponds to the log-phase optimum. Organisms experiencing prolonged periods of balanced growth would therefore be expected to reduce overabundance, which is not observed [1]. Furthermore, the environmental fluctuation model explains overabundance for proteins related to metabolic

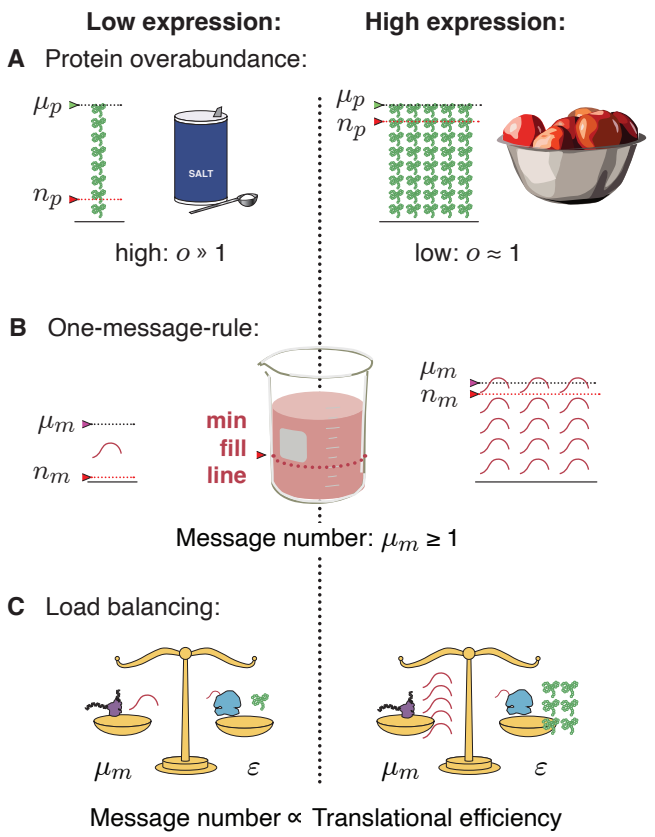

FIG. 9. **Central dogma regulatory principles. Panel A: Overabundance.** Low-expression essential genes are expressed with high overabundance; whereas, high-expression essential genes are expressed with low overabundance. Pie-baking analogy: The low cost of essential low-quantity ingredients (*e.g.* salt) makes it optimal to buy them in great excess, while the higher cost of essential large-quantity ingredients (*e.g.* apples) makes it optimal to buy only what is needed. **Panel B: One-message-rule.** Robust expression of essential genes requires them to be transcribed above a threshold of one message per cell cycle. **Panel C: Load balancing.** In eukaryotic cells, optimal fitness is achieved by balancing transcription and translation: The optimal message number is proportional to the optimal translation efficiency. High (low) expression levels are achieved by high (low) levels of transcription followed by high (low) levels of translation per message.

processes, whereas the RLTO model predicts overabundance generically, dependent only on message number. This generic overabundance appears to be much more consistent with observations (*e.g.* [16]).

Load balancing. A second non-trivial prediction of the RLTO Model is that translation efficiency and message number should be roughly proportional. Qualitatively, this strategy allows expression levels to be increased while distributing the added metabolic load between transcription, which reduces noise, and translation, which does not affect the noise. We predict the optimal translation efficiency versus message number, which matches the observations in eukaryotic cells (Fig. 7BC). However, in *E. coli*, the translation efficiency and message number are not strongly correlated (Fig. 7D). Why does this organism appear not to load balance? In the supplementary material, we demonstrate that the observed translation efficiency is consistent with the RLTO model, augmented by a ribosome-per-message limit. Hausser *et al.* have proposed just such a limit, based on the ribosome footprint on mRNA molecules [14]. (See Supplementary Material Sec. A 9.) Although this augmented model is consistent with central dogma regulation in *E. coli*, it is not a complete rationale. This proposed translation-rate limit could be circumvented by increasing the lifetime of *E. coli* messages, which would increase the translation efficiency. Why the message lifetime is as short as observed will require a more detailed *E. coli*-specific analysis.

Comparisons to previous work. Our analysis is not the first to consider the trade-off between noise and metabolic load. Notably, Hausser *et al.* performed a more limited analysis [14]; however, their model does not share the rich phenomenology we report. What is the difference between these two analyses?

Hausser *et al.* assume a symmetric (not an asymmetric) fitness landscape and consider only the metabolic cost of transcription (but not translation). Their model depends on two (not one) gene-specific parameters: an optimal protein number and a sensitivity, which defines the curvature of the fitness [14]. The authors maximize fitness with respect to the transcription rate (but not the translation rate) and the condition they derive depends on the two (not one) unknown, gene-specific parameters. As a result, this condition is not predictive of global regulatory trends without non-trivial, gene-specific measurements or assumptions about the unknown sensitivity. (See Supplementary Material Sec. E.)

Implications of noise. What are the biological implications of gene expression noise? Many important proposals have been made, including bet-hedging strategies, the necessity of feedback in gene regulatory networks, *etc.* [11]. Our model suggests that robustness to noise fundamentally shapes the central dogma regulatory program. With respect to message number, the one-message-rule sets a lower bound on the transcription rate of essential genes. (See Fig. 9B.) With respect to protein expression, robustness to noise has two important implications: Protein overabundance significantly increases protein levels above what would be required in the absence of noise and therefore reshapes the metabolic budget. (See Fig. 9A.) Robustness to noise also gives rise to load balancing, the proportionality of the optimal transcription and translation rates. (See Fig. 9C.) Not only does robustness to noise affect central dogma regulation, but there is an important reciprocal effect: Load balancing changes the global scaling relation between noise and protein abundance. (See Fig. 8B.)

The principles that govern central dogma regulation.

In summary, the Robustness-Load Trade-Off (RLTO) model describes a number of key regulatory principles that predict the function of the central dogma and are summarized in Fig. 9. For high-expression genes, load balancing implies that gene expression consists of both high-amplification translation and transcription. The resulting expression level has low overabundance relative to the threshold required for function. In contrast, for essential low-expression genes, a three-fold strategy is implemented: (i) overabundance raises the mean protein levels far above the threshold required for function, (ii) load balancing and (iii) the one message rule ensure that message number is sufficiently large to lower the noise of inherently-noisy low-expression genes. We anticipate that these regulatory principles, in particular protein overabundance, will have significant impact, not only on our understanding of central dogma regulation specifically, but in understanding the rationale for protein expression level and function in many biological processes.

Data availability. We include a source data file which includes the estimated message numbers as well as essential/nonessential classifications for each organism.

Acknowledgments. The authors would like to thank B. Traxler, A. Nourmohammad, J. Mougous, K. Cutler, M. Cosentino-Lagomarsino, S. van Teeffelen, and S. Murray. This work was supported by NIH grant R01-GM128191.

Author contributions: T.W.L., H.K.J.C., D.H. and P.A.W. conceived the research. T.W.L. and P.A.W. performed the analysis. H.K.J.C. and D.H. performed experiments and analysis. T.W.L., H.K.J.C., D.H. and P.A.W. wrote the paper.

Competing interests: The authors declare no competing interests.

[1] Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**, 588–92 (2005).
[2] Keren, L. *et al.* Massively parallel interrogation of the effects of gene expression levels on fitness. *Cell* **166**, 1282–1294.e18 (2016).
[3] Mori, M., Schink, S., Erickson, D. W., Gerland, U. & Hwa, T. Quantifying the benefit of a proteome reserve in fluctuating environments. *Nat Commun* **8**, 1225 (2017).
[4] Pedersen, S. *Escherichia coli* ribosomes translate in vivo with variable rate. *EMBO J* **3**, 2895–8 (1984).
[5] Young, R. & Bremer, H. Polypeptide-chain-elongation rate in *Escherichia coli* B/r as a function of growth rate. *Biochem J* **160**, 185–94 (1976).
[6] Dennis, P. P. & Bremer, H. Macromolecular composition during steady-state growth of *Escherichia coli* B-r. *J Bacteriol* **119**, 270–81 (1974).
[7] Koch, A. L. Overall controls on the biosynthesis of ribosomes in growing bacteria. *J Theor Biol* **28**, 201–31 (1970).
[8] Klumpp, S., Scott, M., Pedersen, S. & Hwa, T. Molecular crowding limits translation and cell growth. *Proc Natl Acad Sci U S A* **110**, 16754–9 (2013).
[9] Belliveau, N. M. *et al.* Fundamental limits on the rate of bacterial growth and their influence on proteomic composition. *Cell Syst* **12**, 924–944.e2 (2021).
[10] Lambert, G. & Kussell, E. Memory and fitness optimization of bacteria under fluctuating environments. *PLoS Genet* **10**, e1004556 (2014).
[11] Raser, J. M. & O'Shea, E. K. Noise in gene expression: origins, consequences, and control. *Science* **309**, 2010–3 (2005).
[12] Newman, J. R. S. *et al.* Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441**, 840–6 (2006).
[13] Taniguchi, Y. *et al.* Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–8 (2010).
[14] Hausser, J., Mayo, A., Keren, L. & Alon, U. Central dogma rates and the trade-off between precision and economy in gene expression. *Nat Commun* **10**, 68 (2019).
[15] Peters, J. M. *et al.* A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. *Cell* **165**, 1493–1506 (2016).
[16] Gallagher, L. A., Bailey, J. & Manoil, C. Ranking essential bacterial processes by speed of mutant death. *Proc Natl Acad Sci U S A* **117**, 18010–18017 (2020).
[17] Donati, S. *et al.* Multi-omics analysis of crispri-knockdowns identifies mechanisms that buffer decreases of enzymes in e. coli metabolism. *Cell Syst* **12**, 56–67.e6 (2021).
[18] Paulsson, J. & Ehrenberg, M. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Phys Rev Lett* **84**, 5447–50 (2000).
[19] Friedman, N., Cai, L. & Xie, X. S. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys Rev Lett* **97**, 168302 (2006).
[20] Joseph W. Lengeler, H. G. S., Gerhart Drews (ed.) *Biology of the Prokaryotes* (Georg Thieme Verlag, Rüdigerstrasse 14, D-70469 Stuttgart, Germany, 1998).
[21] I, S. J., S., F. J. & L., H. W. *Chemical Kinetics and Dynamics* (Prentice-Hall, 1999), 2nd edn.
[22] Choi, H. K. J., ... & Wiggins, P. A. Tba (2023).
[23] Huang, D., Lo, T., Merrikh, H. & Wiggins, P. A. Characterizing stochastic cell-cycle dynamics in exponential growth. *Phys Rev E* **105**, 014420 (2022).
[24] Balakrishnan, R. *et al.* Principles of gene regulation quantitatively connect DNA to RNA and proteins in bacteria. *Science* **378**, eabk2066 (2022).
[25] Silvis, M. R. *et al.* Morphological and transcriptional responses to crispri knockdown of essential genes in escherichia coli. *mBio* **12**, e0256121 (2021).
[26] Kafri, M., Metzl-Raz, E., Jona, G. & Barkai, N. The cost of protein production. *Cell Rep* **14**, 22–31 (2016).
[27] Metzl-Raz, E., Kafri, M., Yaakov, G. & Barkai, N. Gene transcription as a limiting factor in protein production

and cell growth. *G3 (Bethesda)* **10**, 3229–3242 (2020).
[28] Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–41 (2003).
[29] Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–42 (2011).
[30] Bar-Even, A. *et al.* Noise in protein expression scales with natural protein abundance. *Nat Genet* **38**, 636–43 (2006).
[31] Bosch, J. A., Chen, C.-L. & Perrimon, N. Proximity-dependent labeling methods for proteomic profiling in living cells: An update. *WIREs Developmental Biology* **10** (2020). URL https://doi.org/10.1002/wdev.392.
[32] Monod, J., Pappenheimer, A. M., Jr & Cohen-Bazire, G. The kinetics of the biosynthesis of beta-galactosidase in *Escherichia coli* as a function of growth. *Biochim Biophys Acta* **9**, 648–60 (1952).
[33] Novick, A. & Weiner, M. Enzyme induction as an all-or-none phenomenon. *Proc Natl Acad Sci U S A* **43**, 553–66 (1957).
[34] Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–3 (1970).
[35] Hargrove, J. L. & Schmidt, F. H. The role of mRNA and protein stability in gene expression. *FASEB J* **3**, 2360–70 (1989).
[36] Koch, A. L. & Levy, H. R. Protein turnover in growing cultures of *Escherichia coli*. *J Biol Chem* **217**, 947–57 (1955).
[37] Martin-Perez, M. & Villén, J. Determinants and regulation of protein turnover in yeast. *Cell Syst* **5**, 283–294.e5 (2017).
[38] Gillespie, D. A rigorous derivation of the chemical master equation. *Physica A* **188**, 404–425 (1992).
[39] Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* **81**, 2340–2361 (1977).
[40] Bernstein, J. A., Khodursky, A. B., Lin, P.-H., Lin-Chao, S. & Cohen, S. N. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci U S A* **99**, 9697–702 (2002).
[41] Chen, H., Shiroguchi, K., Ge, H. & Xie, X. S. Genome-wide study of mRNA degradation and transcript elongation in *Escherichia coli*. *Mol Syst Biol* **11**, 808 (2015).
[42] (2023). URL https://en.wikipedia.org/wiki/Gamma_distribution.
[43] Huang, D., Lo, T., Merrikh, H. & Wiggins, P. A. Quantitative in vivo measurement of fork velocity in bacterial cells by deep sequencing. *Nature Communications (In press.)* (2023).
[44] Wang, P. *et al.* Robust growth of escherichia coli. *Curr Biol* **20**, 1099–103 (2010).
[45] Baba, T., Huan, H.-C., Datsenko, K., Wanner, B. L. & Mori, H. The applications of systematic in-frame, single-gene knockout mutant collection of *Escherichia coli* K-12. *Methods Mol Biol* **416**, 183–94 (2008).
[46] Gerdes, S. Y. *et al.* Experimental determination and system level analysis of essential genes in *Escherichia coli* mg1655. *J Bacteriol* **185**, 5673–84 (2003).
[47] Goodall, E. C. A. *et al.* The essential genome of *Escherichia* coli k-12. *mBio* **9** (2018).
[48] Mehta, P., Casjens, S. & Krishnaswamy, S. Analysis of the lambdoid prophage element e14 in the *E. coli* k-12 genome. *BMC Microbiol* **4**, 4 (2004).
[49] Gerdes, S. Y. *et al.* Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J Bacteriol* **185**, 5673–84 (2003).
[50] Arfken, G. *Mathematical Methods for Physicists* (Academic Press, Inc., San Diego, 1985), third edn.
[51] Milo, R., Jorgensen, P., Moran, U., Weber, G. & Springer, M. Bionumbers–the database of key numbers in molecular and cell biology. *Nucleic Acids Res* **38**, D750–3 (2010).
[52] Bartholomäus, A. *et al.* Bacteria differently regulate mRNA abundance to specifically respond to various stresses. *Philos Trans A Math Phys Eng Sci* **374** (2016).
[53] Milo, R. What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays* **35**, 1050–5 (2013).
[54] Chia, L. L. & McLaughlin, C. The half-life of mRNA in *Saccharomyces cerevisiae*. *Mol Gen Genet* **170**, 137–44 (1979).
[55] Blevins, W. R. *et al.* Extensive post-transcriptional buffering of gene expression in the response to severe oxidative stress in baker's yeast. *Sci Rep* **9**, 11005 (2019).
[56] Hereford, L. M. & Rosbash, M. Number and distribution of polyadenylated RNA sequences in yeast. *Cell* **10**, 453–62 (1977).
[57] von der Haar, T. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC Syst Biol* **2**, 87 (2008).
[58] Zenklusen, D., Larson, D. R. & Singer, R. H. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol* **15**, 1263–71 (2008).
[59] Pelechano, V., Chávez, S. & Pérez-Ortín, J. E. A complete set of nascent transcription rates for yeast genes. *PLoS One* **5**, e15442 (2010).
[60] Miura, F. *et al.* Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs. *BMC Genomics* **9**, 574 (2008).
[61] Alberts, B. *et al.* *Molecular Biology of the Cell* (Garland, 2002), 4th edn.
[62] Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S. & Garrels, J. I. A sampling of the yeast proteome. *Mol Cell Biol* **19**, 7357–68 (1999).
[63] van Leeuwen, J. *et al.* Systematic analysis of bypass suppression of essential genes. *Mol Syst Biol* **16**, e9828 (2020).
[64] de Godoy, L. M. F. *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–4 (2008).
[65] Yang, E. *et al.* Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res* **13**, 1863–72 (2003).
[66] Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
[67] Velculescu, V. E. *et al.* Analysis of human transcriptomes. *Nat Genet* **23**, 387–8 (1999).
[68] Cooper, G. M. *The Cell: A Molecular Approach. 2nd edition* (Sinauer Associates 2000, 2000).
[69] Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–101 (2015).
[70] Lahtvee, P.-J. *et al.* Absolute quantification of protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in yeast. *Cell Syst* **4**, 495–504.e5 (2017).
[71] Hellton, K. H. & Thoresen, M. The impact of measurement error on principal component analysis. *Scandinavian Journal of Statistics* **41**, 1051–1063 (2014). URL https://doi.org/10.1111/sjos.12083.
[72] Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–6 (2002).

[73] Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci U S A* **99**, 12795–800 (2002).

CONTENTS

Introduction 1

Results 2

Discussion 11

References 13

A. Supplemental analysis of the RLTO model 17
1. Detailed description of the noise model 17
a. Stochastic kinetic model for central dogma. 17
b. Statistical model for protein abundance. 17
c. Gillespie simulation of the telegraph model 17
d. Gamma function and distribution conventions 18
2. The derivation of the RLTO growth rate 18
a. Metabolic load 18
b. Growth rate with arrest 18
c. RLTO growth rate 19
d. When the ln approximation is avoided... 19
e. Single-gene equation 19
3. Message number and translation efficiency optimization 19
4. Modifications of the RLTO model for bacterial cells 20
a. Optimization of message number only 21
b. Adaptation of the RLTO model to a noise floor 21
5. Arrest probability 22
6. Discussion of *E. coli* essential genes below the one-message-rule threshold 22
7. Measurements of the load ratio 23
8. Increased protein load analysis 23
a. Derivation 23
b. Increased protein cost reduces the optimal translation efficiency. 23
9. Analysis of translational limits. 24
10. Parameter values for figures 25
11. Estimate of the message cost and metabolic load 25
a. Detailed protocol 25
b. Estimate the message cost and metabolic load in yeast 26
c. Estimate the message cost and metabolic load in human cells 26
12. Prediction of the proteome fraction 26
a. RLTO: proteome fraction 26
b. Constant-translation-efficiency model: proteome fraction 27
c. Sources of experimental data for proteome fraction analysis 27

B. Analysis of alternative models 27
1. Threshold (RLTO) model 27
2. Model 2: Slow-Growth model 27
3. Model 3: Symmetric model 27
4. Conclusions from fitness-landscape analysis 28

C. Quantitation of central dogma parameters for one-message-rule 28
1. Selection of central dogma parameter estimates 28
a. *E. coli* data 28
b. Yeast data 29
c. Human data 29
2. Quantitative estimates of central dogma parameters 29
a. Estimating the *cellular message number*: $\mu_{m/c}$ 29
b. Estimating the *transcription rate*: β_m 29
c. Estimating the *message number*: μ_m 29
d. Histograms 30

D. Supplemental analysis of Noise-Protein-Abundance Relation in Yeast 30
1. Estimating *protein number* (μ_p) for the noise analysis 30
2. Empirical models for yeast gene expression 30
a. The meaning of the error estimates 30
b. Empirical model for protein number 31
c. Empirical model for message number 31
d. Empirical model for translation efficiency 31
e. Empirical model for the coefficient of variation 32
3. Supplemental analysis of gene expression noise 32
a. Inclusion of noise floor in the yeast analysis 32
b. Hypothesis test I 33
c. Hypothesis test II 33
d. Maximum likelihood estimates of the parameters 33
e. Details: Statistical details MLE estimate of the variance 33

E. Comments on the Hausser *et al.* analysis 33

Appendix A: Supplemental analysis of the RLTO model

1. Detailed description of the noise model

a. *Stochastic kinetic model for central dogma.*

The telegraph model for the central dogma describes multiple steps in the gene expression process: Transcription generates mRNA messages [34]. These messages are then translated to synthesize the protein gene products [34]. Both mRNA and protein are subject to degradation and dilution [35]. At the single cell level, each of these processes are stochastic. We will model these processes with the stochastic kinetic scheme [34]:

$$\begin{array}{ccccc} \text{DNA} & \xrightarrow{\beta_m} & \text{mRNA} & \xrightarrow{\beta_p} & \text{Protein} \\ & & \gamma_m \big\downarrow & & \gamma_p \big\downarrow \\ & & \emptyset & & \emptyset, \end{array} \tag{A1}$$

where β_m is the transcription rate (s^{-1}), β_p is the translation rate (s^{-1}), γ_m is the message degradation rate (s^{-1}), and γ_p is the protein effective degradation rate (s^{-1}). The message lifetime is $T_m \equiv \gamma_m^{-1}$. For most proteins in the context of rapid growth, dilution is the dominant mechanism of protein depletion and therefore γ_p is approximately the growth rate [13, 36, 37]: $\gamma_p = T_{cc}^{-1} \ln 2$, where T_{cc} is the doubling time.

b. *Statistical model for protein abundance.*

To study the stochastic dynamics of gene expression, we used a stochastic Gillespie simulation [38, 39]. (See Supplemental Material Sec. A 1 c.) In particular, we were interested in the explicit relation between the kinetic parameters $(\beta_m, \gamma_m, \beta_p, \gamma_p)$ and experimental observables.

Consistent with previous reports [18, 19], we find that the distribution of protein number per cell (at cell birth) was described by a gamma distribution: $N_p \sim \Gamma(a, \theta)$, where N_p is the protein number at cell birth and Γ is the gamma distribution which is parameterized by a scale parameter θ and a shape parameter a. (See Supplementary Material Sec. A 1 d.) The relation between the four kinetic parameters and these two statistical parameters has already been reported, and have clear biological interpretations [19]: The scale parameter:

$$\theta = \varepsilon \ln 2, \tag{A2}$$

is proportional to the translation efficiency:

$$\varepsilon \equiv \frac{\beta_p}{\gamma_m}, \tag{A3}$$

where β_p is the translation rate and γ_m is the message degradation rate. ε is understood as the mean number of proteins translated from each message transcribed. The shape parameter a can also be expressed in terms of the kinetic parameters [19]:

$$a = \frac{\beta_m}{\gamma_p}; \tag{A4}$$

however, we will find it more convenient to express the scale parameter in terms of the cell-cycle message number:

$$\mu_m \equiv \beta_m T_{cc} = a \ln 2, \tag{A5}$$

which can be interpreted as the mean number of messages transcribed per cell cycle. Forthwith, we will abbreviate this quantity *message number* in the interest of brevity.

c. *Gillespie simulation of the telegraph model*

Protein distributions of the telegraph model for *E. coli* were simulated with a Gillespie algorithm. Assuming the lifetime of the cell cycle ($T_{cc} = 30$ min) [40], mRNA lifetime ($T_m = 2.5$ min) [41], and translation rate ($\beta_p \approx 500$ hr^{-1}), the protein distributions for several mean expression levels were numerically generated for exponential growth with 100,000 stochastic cell divisions, with protein partitioned at division following the binomial distribution.

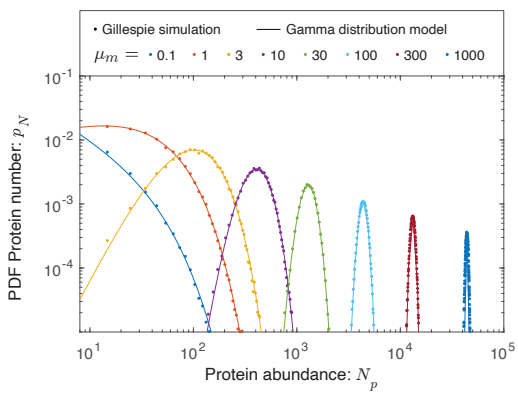

FIG. 10. **The protein abundance is approximately gamma distributed.** Protein abundance was modeled for eight different transcription rates using a Gillespie simulation, including the stochastic partitioning of the proteins between daughter cells at cell division. The range in abundance matches the observed range of expression levels in the cell. We observed that the simulated protein abundances were well fit by gamma distributions.

The gamma distributions for each mean message number with scale and shape parameters determined by the corresponding translation efficiency and message number ($\theta = \varepsilon \ln 2$, $a = \frac{\mu_m}{\ln 2}$) as used for the Gillespie simulation were also plotted with the protein distributions. We observe an excellent match between these Gillespie simulations and the statistical noise model (*i.e.* gamma function) as shown in Fig. 10.

d. *Gamma function and distribution conventions*

There are a number of conflicting conventions for the gamma function and distribution arguments. We will use those defined on Wikipedia [42]. The gamma distributed random variable X will be written:

$$X \sim \Gamma(a, \theta), \tag{A6}$$

where a is the shape parameter and θ is the scale parameter. The PDF of the distribution is:

$$p_\Gamma(x|a,\theta) \equiv \frac{x^{a-1}}{\theta^a \Gamma(a)} e^{-x/\theta}, \tag{A7}$$

where $\Gamma(a)$ is the gamma function. The CDF is therefore:

$$\begin{aligned} P_\Gamma(x|a,\theta) &= \int_0^x \mathrm{d}x'\, p_\Gamma(x'|a,\theta), && \text{(A8)} \\ &= P_\Gamma(\tfrac{x}{\theta}|a,1), && \text{(A9)} \\ &= \int_0^{x/\theta} \mathrm{d}x'\, \frac{x^{a-1}}{\Gamma(a)} e^{-x}, && \text{(A10)} \\ &= \gamma(a, x/\theta), && \text{(A11)} \end{aligned}$$

where γ is the regularized incomplete gamma function.

2. The derivation of the RLTO growth rate

a. *Metabolic load*

In the absence of cell cycle arrest, we will assume that the cell cycle duration τ is proportional to the overall metabolic load of the messages, proteins, and other cellular components [20]. Focusing on a particular gene:

$$\tfrac{\tau}{\delta\tau} = N_0 + \lambda\mu_m + \mu_p, \tag{A12}$$

where $\delta\tau$ is a constant with units of time, N_0 is the metabolic load of everything but gene i (in units of protein equivalents), λ is the message cost, the metabolic load associated with an mRNA molecule relative to a single protein molecule of the gene product. We will assume the parameters N_0, $\delta\tau$ and λ are global (*i.e.* gene independent), while the message number μ_m and translation efficiency ε, and therefore the protein number (Eq. 1) are gene specific. The total metabolic load in protein equivalents is:

$$N_0 \equiv L_0 + \sum_i \{(\lambda + \varepsilon)\mu_m\}_i, \tag{A13}$$

where L_0 is the load of non-protein and message cellular components and the sum runs over all genes. The metabolic load of transcription for gene i is $\lambda\mu_{m,i}$ and for translation $\mu_{p,i} = \varepsilon_i \mu_{m,i}$.

Although the global parameters N_0, $\delta\tau$ and λ provide an intuitive representation of the model, the relative growth rate depends on fewer parameters. Let k and k_0 be the growth rates in the presence and absence of the metabolic load of gene i. The relative growth rate is:

$$\tfrac{k_0}{k} = 1 + (\Lambda + E)\mu_m, \tag{A14}$$

where we have introduced two new reduced parameters: the relative load, defined as $\Lambda \equiv \lambda/N_0$, represents the ratio of the metabolic load of a single message to the total load and the relative translation efficiency, defined $E \equiv \varepsilon/N_0$, which is the ratio of the number of proteins translated per message to the total metabolic load N_0. If we neglect the difference between the total metabolic load and the number of proteins, the proteome fraction for gene i is $\Phi_p = E\mu_m$. Both reduced parameters, Λ and E are extremely small. In *E. coli*, we estimate that both Λ and E are roughly 10^{-5} and they are smaller still for eukaryotic cells. (See Supplementary Material Sec. A 11.)

b. *Growth rate with arrest*

For completeness, we provide a derivation of the growth rate with stochastic cell-cycle arrest that we have previously described [23]. Starting from the exponential mean expression for the population growth rate [23]:

$$k = \tfrac{\ln 2}{T}, \tag{A15}$$

where k is the population growth rate and

$$\overline{T} \equiv -\frac{1}{k} \ln \mathbb{E}_T \exp(-kT), \tag{A16}$$

is the exponential mean where T is the stochastic cell cycle duration [23, 43].

Let P_+ be the probability of growth. When the cells are growing, the cell cycle duration is given by the metabolic load cell-cycle duration τ (Eq. A12). Evaluating the expectation gives:

$$\ln 2 = -\ln P_+ + k\tau, \tag{A17}$$

which can be rearranged to give an expression for the population growth rate k:

$$k = \tau^{-1} \ln 2P_+. \tag{A18}$$

As expected, the growth rate goes down as the probability of growth P_+ decreases, stopping completely at $P_+ = \frac{1}{2}$.

c. RLTO growth rate

In the RLTO model, we will assume the probability of growth is the probability that all essential protein numbers are above threshold. We will further assume that each protein number is independent, and therefore:

$$P_+ = \prod_{i \in \mathscr{E}} \Pr\{N_p > n_p\}_i, \tag{A19}$$

where $\mathscr{E}$ is the set of essential genes. Clearly, this assumption of independence fails in the context of polycistronic messages. We will discuss the significance of this feature of bacterial cells elsewhere, but we will ignore it in the current context.

As we will discuss, the probability of arrest of any protein i to be above threshold is extremely small. It is therefore convenient to work in terms of the CDFs which are very close to zero:

$$\begin{aligned} \ln P_+ &= \sum_{i \in \mathscr{E}} \ln\left(1 - \Pr\{N_p < n_p\}_i\right), && \text{(A20)} \\ &\approx -\sum_{i \in \mathscr{E}} \Pr\{N_p < n_p\}_i\}, && \text{(A21)} \\ &= -\sum_{i \in \mathscr{E}} \left[\gamma(\tfrac{\mu_m}{\ln 2}, \tfrac{n_p}{\varepsilon \ln 2})\right]_i, && \text{(A22)} \end{aligned}$$

where γ is the regularized incomplete gamma function and the CDF of the gamma distribution. Finally, we will be interested in the quantity:

$$\ln \ln 2P_+ \approx \ln \ln 2 + -\sum_{i \in \mathscr{E}} \tfrac{1}{\ln 2} \left[\gamma(\tfrac{\mu_m}{\ln 2}, \tfrac{n_p}{\varepsilon \ln 2})\right]_i, \tag{A23}$$

where we have again expanded the natural logarithm in the limit that the CDF is small.

d. When the ln approximation is avoided...

The approximation discussed in the previous section is extremely well justified at the optimal central dogma parameters; however, there are a set of figures where we cannot use it. In the fitness landscape figures (Figs. 3 and 11), we compute the fitness not just at the optimal values but far from them. Here we cannot approximate the natural log and we use the full expression in Eq. A20.

e. Single-gene equation

By combining Eqs. A18, A12 and A23, we can write an expression for the growth rate:

$$\begin{aligned} \ln k &= -\ln\left(L_0 + \sum_i \left[(\lambda + \varepsilon)\mu_m\right]_i\right) + \ldots \\ &\quad + \ln \ln 2 - \sum_{i \in \mathscr{E}} \tfrac{1}{\ln 2} \left[\gamma(\tfrac{\mu_m}{\ln 2}, \tfrac{n_p}{\varepsilon \ln 2})\right]_i. \end{aligned} \tag{A24}$$

We will work in the large multiplicity limit where each of the arrest probabilities is very small. To optimize the expression of gene i we can drop all constant terms and write the simplified expression:

$$\begin{aligned} \ln k &= -\ln\left(N_0' + (\lambda + \varepsilon)\mu_m\right) + \ldots \\ &\quad + \ln \ln 2 - \tfrac{1}{\ln 2}\gamma(\tfrac{\mu_m}{\ln 2}, \tfrac{n_p}{\varepsilon \ln 2}) + \ldots, \end{aligned} \tag{A25}$$

where we have made the essential gene i subscript implicit and we consider a total load minus gene i:

$$N_0' \equiv L_0 + \sum_{j \neq i} \left[(\lambda + \varepsilon)\mu_m\right]_j, \tag{A26}$$

interpreted as the metabolic load of all genes but i. However, we will work in the large multiplicity limit where we ignore the distinction between N_0 and N_0'. The fitness landscape for different gene expression parameters is shown from four different perspective in Fig. 11.

3. Message number and translation efficiency optimization

Starting with Eq. 4 for the growth rate, we set the partial derivative with respect to message number equal to zero:

$$0 = -\tfrac{\lambda + \hat{\varepsilon}}{N_0} - \tfrac{1}{(\ln 2)^2}\gamma_{,1}(\tfrac{\hat{\mu}_m}{\ln 2}, \tfrac{n_p}{\hat{\varepsilon} \ln 2}), \tag{A27}$$

where we use the canonical comma notation to show which argument of γ has been differentiated. Next we differentiate with respect to the translation efficiency to generate a second optimization condition:

$$0 = -\tfrac{\hat{\mu}_m}{N_0} + \tfrac{n_p}{\hat{\varepsilon}^2 (\ln 2)^2}\gamma_{,2}(\tfrac{\hat{\mu}_m}{\ln 2}, \tfrac{n_p}{\hat{\varepsilon} \ln 2}). \tag{A28}$$

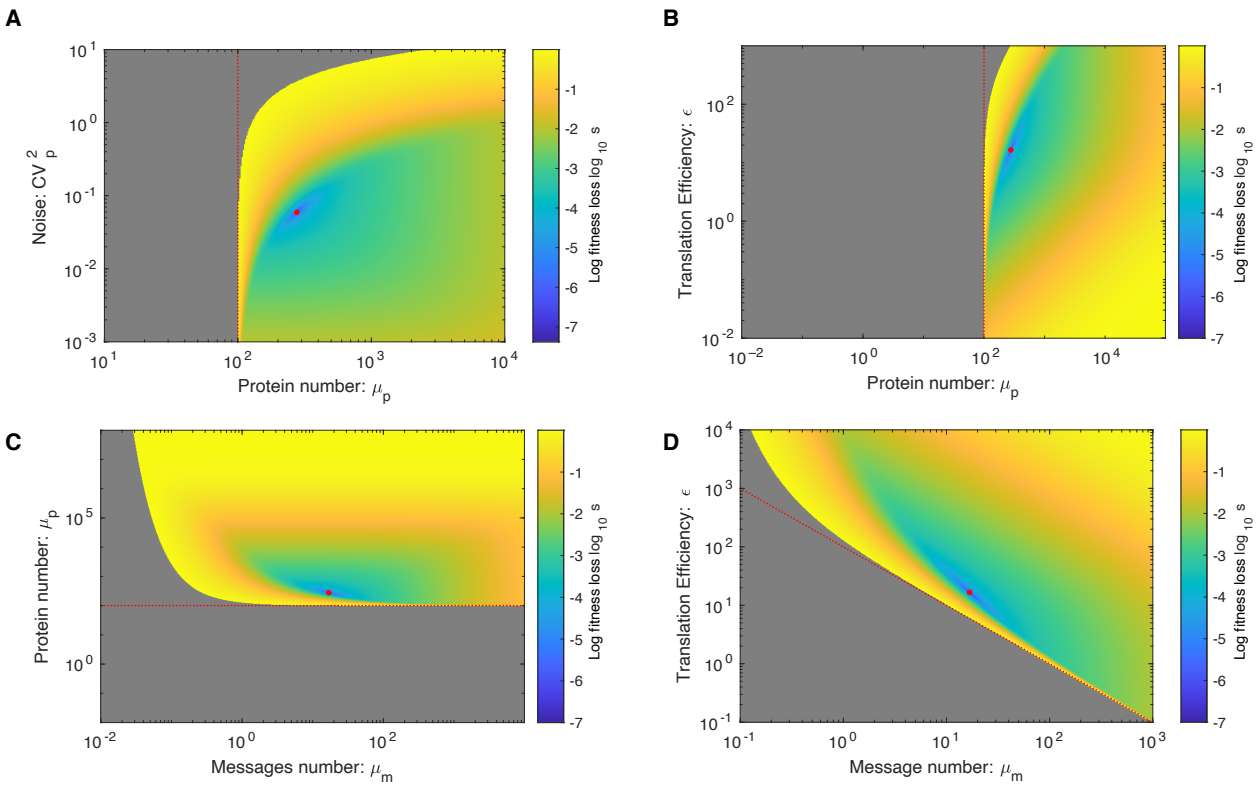

FIG. 11. **Four perspectives on the fitness landscape.** In each landscape, the red circle represents the fitness optimum. The red dotted line represents the mean protein number equal to the protein threshold $n_p = 10^2$. Here fitness is quantified by the log growth rate: $s = \log k_{\max}/k$. **Panel A: Mean protein number versus noise. Panel B: Mean protein number versus translation efficiency. Panel C: Mean message versus protein numbers. Panel D: Message number versus translation efficiency.**

We will work in the large multiplicity limit where the overall metabolic load is much smaller than the metabolic load associated with any single gene: $N_0 \gg (\lambda + \hat{\varepsilon})\hat{\mu}_m$. Next, we eliminate the threshold n_p in favor of the optimal overabundance:

$$\hat{o} \equiv \frac{\hat{\mu}_p}{n_p} = \frac{\hat{\varepsilon}\hat{\mu}_m}{n_p}, \tag{A29}$$

in both Eqs. A27 and A28. Eq.A28 can now be solved for the optimal translation effeciency:

$$\hat{\varepsilon} = \frac{N_0}{\hat{o}(\ln 2)^2}\gamma_{,2}\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}\right). \tag{A30}$$

If we reinterpret γ as the CDF of the gamma distribution, we can rewrite this equation in terms of the gamma distribution PDF:

$$\hat{\varepsilon} = \frac{N_0}{\ln 2} p_\Gamma\left(\mu_m \middle| \frac{\hat{\mu}_m}{\ln 2}, \hat{o}\ln 2\right), \tag{A31}$$

which will be the optimization equation for the translation efficiency.

To derive the optimization condition for the message number μ_m, we substitute Eq. A30 into Eq. A27:

$$\frac{\lambda \ln 2}{N_0} = -\frac{1}{\hat{o}\ln 2}\gamma_{,2}\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}\right) - \frac{1}{\ln 2}\gamma_{,1}\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}\right). \tag{A32}$$

The two terms on the RHS can now be collected as the single partial derivative of message number μ_m:

$$\Lambda \ln 2 = -\partial_{\hat{\mu}_m}\gamma\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}\right), \tag{A33}$$

where the relative load is $\Lambda \equiv \lambda/N_0$.

The two optimization equations are summarized below:

$$\Lambda \ln 2 = -\partial_{\hat{\mu}_m}\gamma\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}\right), \tag{A34}$$

$$\frac{\hat{E}}{\Lambda} = \frac{\hat{\varepsilon}}{\lambda} = \frac{1}{\Lambda \ln 2} p_\Gamma\left(\hat{\mu}_m \middle| \frac{\hat{\mu}_m}{\ln 2}, \hat{o}\ln 2\right), \tag{A35}$$

where despite the factor of Λ in the denominator of the RHS of Eq. A35, the RHS depends very weakly on Λ. See Fig. 7A. The overabundance is shown for a range of relative loads in Fig. 12. The optimal translation efficiency and scaled translation efficiency are shown for a range of relative loads in Fig. 13.

4. Modifications of the RLTO model for bacterial cells

There are two features of the bacterial cell that are distinct relative to eukaryotes: (i) constant translation efficiency and (ii) a high noise floor. We will consider both.

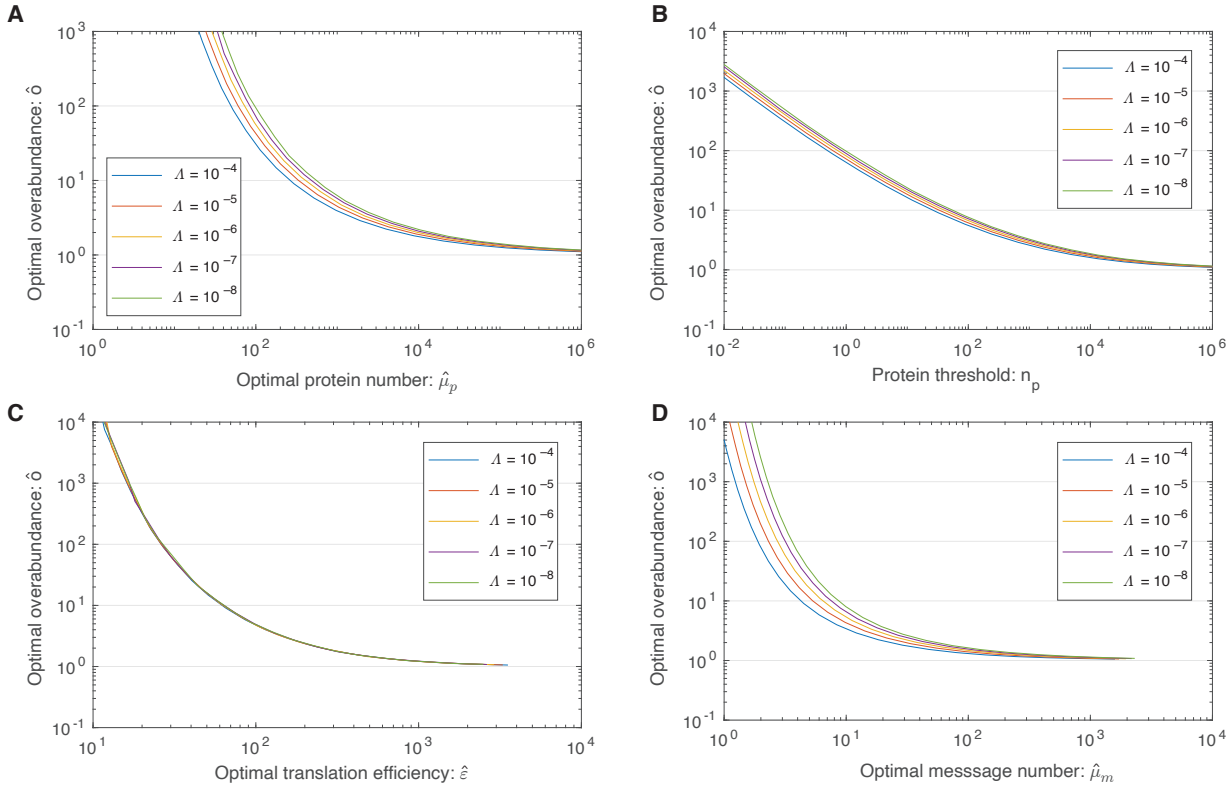

FIG. 12. **Four perspectives on the dependence of optimal overabundance on relative load Λ.** All these calculations are performed for protein cost $\lambda = 100$ in order to give real numbers in molecules per cell. **Panel A: Overabundance as a function of protein number.** Overabundance decreases as protein number increases. These calculations are λ *dependent*. **Panel B: Overabundance as a function of the protein threshold.** Overabundance decreases as protein threshold increases. These calculations are λ *dependent*. **Panel C: Overabundance as a function of translation efficiency.** Overabundance decreases as translation efficiency increases. These calculations are λ *dependent*. **Panel D: Overabundance as a function of message number.** Overabundance decreases as message number increases. These calculations are λ *independent*.

a. *Optimization of message number only*

Consider the special case of optimizing the message number only at fixed translation efficiency. Eq. A27 is essentially the condition; however, in this case it makes sense to adsorb both the message and protein metabolic load into a single metabolic load. The optimum message number satisfies the equation:

$$\frac{(\lambda+\varepsilon)\ln 2}{N_0} = -[\partial_{\hat{\mu}_m}\gamma(\hat{\mu}_m, \hat{n}_m)]_{\hat{n}_m = \frac{\hat{\mu}_m}{\hat{o}}}. \tag{A36}$$

We define a modified relative load:

$$\Lambda' \equiv \frac{(\lambda+\varepsilon)}{N_0}, \tag{A37}$$

and substitute this into the optimum message number equation:

$$\Lambda' \ln 2 = -[\partial_{\hat{\mu}_m}\gamma(\hat{\mu}_m, \hat{n}_m)]_{\hat{n}_m = \frac{\hat{\mu}_m}{\hat{o}}}. \tag{A38}$$

which is clearly closely related to Eq. A33.

We compare this modified expression to the original for optimum overabundance as a function of message number in Fig. 4A and demonstrate that the two make nearly identical predictions.

b. *Adaptation of the RLTO model to a noise floor*

In bacterial cells, the noise is dominated by the noise floor for higher expression levels. Including the noise floor, the coefficient of variation squared is [13]:

$$\mathrm{CV}_p^2 = \tilde{a}(\mu_m)^{-1} = \frac{\ln 2}{\mu_m} + C_0, \tag{A39}$$

where $C_0 = 0.1$ for bacterial cells [13]. In spite of the addition of noise from the noise floor, the observed distribution of protein number is still well described by the Gamma distribution [13]; however, we need to modify the statistical parameters to account for the noise floor. The modified gamma parameters are:

$$a = \tilde{a}, \tag{A40}$$

$$\theta = \varepsilon\frac{\mu_m}{\tilde{a}}, \tag{A41}$$

chosen such that the noise is determined by Eq. A39 but the protein number remains:

$$\mu_p = \varepsilon\mu_m, \tag{A42}$$

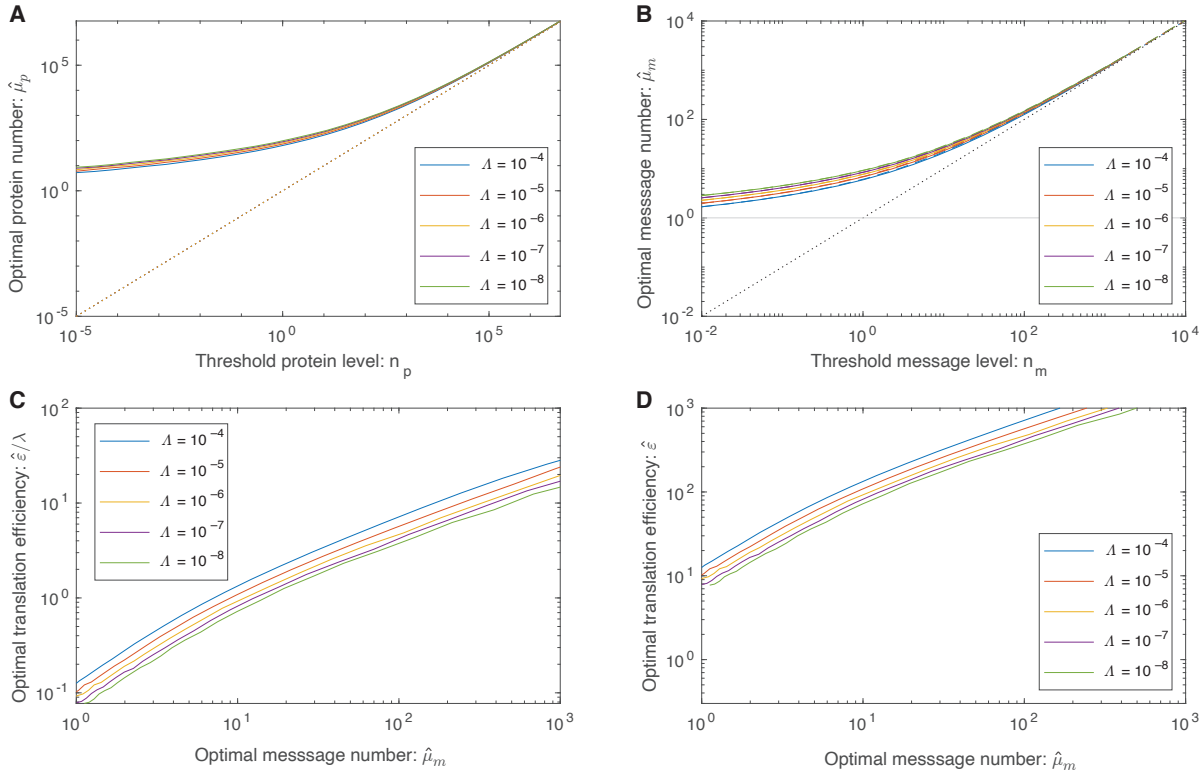

FIG. 13. All these calculations are performed for protein cost $\lambda = 100$ in order to give real numbers in molecules per cell. **Panel A: Protein number versus protein threshold.** At high expression levels, the protein number tracks the protein threshold; however, the one message rule forces the protein number to threshold for low expression levels. These calculations are λ *dependent*. **Panel B: Message number versus message threshold.** At high expression levels, the message number tracks the message threshold; however, the one message rule forces the message number to a threshold close to $\mu_m = 1$ for low expression levels. These calculations are λ *independent*. **Panel C & D: Message number versus translation efficiency.** The optimal translation efficiency grows almost linearly with the optimal message number. The scaled translation efficiency ($\hat{\varepsilon}/\lambda$) is independent of λ while the translation efficiency ($\hat{\varepsilon}$) is dependent on λ.

the product of the message number and translation efficiency.

The qualitative effect of the noise floor is to increase the noise, especially for low-copy messages. Above $\mu_m = 7$ messages, the noise is dominated by the noise floor. Increases in expression above this point have little effect on reducing the noise. As a consequence, the overabundance stays high, even for high copy messages. We compare this modified expression to the original for optimum overabundance as a function of message number in Fig. 4A and demonstrate that bacterial cells are predicted to have much higher overabundance at high expression levels.

5. Arrest probability

The gene i arrest probability at optimal expression is:

$$P_- = \gamma\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\delta \ln 2}\right), \tag{A43}$$

which is independent of the message cost λ. As expected, the arrest probability is extremely low. (See Fig. 15.)

A more interesting experimental quantity is the total arrest probability. We can estimate the total arrest probability in *E. coli* by summing up the arrest probability for all essential genes. The estimated total arrest probability is $P_-^{\text{tot}} = 8.5 \times 10^{-4}$. This result is consistent with single-cell analysis of cell growth where the filamentation rate of MG1655 is roughly 1% per generation, but smaller in the B/r strain [44].

6. Discussion of *E. coli* essential genes below the one-message-rule threshold

Since our own preferred model system is *E. coli*, we focus here. Our essential gene classification was based on the construction of the Keio knockout library [45]. By this classification, 10 essential genes were below threshold. (See Supplementary Material Tab. II.) Our first step

was to determine what fraction of these genes were also classified as essential using transposon-based mutagenesis [46, 47]. Of the 10 initial candidates, only one gene, *ymfK*, was consistently classified as an essential gene in all three studies, and we estimate that its message number is just below the threshold ($\mu_m = 0.4$). *ymfK* is located in the lambdoid prophage element e14 and is annotated as a CI-like repressor which regulates lysis-lysogeny decision [48]. In λ phase, the CI repressor represses lytic genes to maintain the lysogenic state. A conserved function for *ymfK* is consistent with it being classified as essential, since its regulation would prevent cell lysis. However, since *ymfK* is a prophage gene, not a host gene, it is not clear that its expression should optimize host fitness, potentially at the expense of phage fitness. In summary, closer inspection of below-threshold essential genes supports the threshold hypothesis.

7. Measurements of the load ratio

Unfortunately there is somewhat limited data to which to compare the model. The best source we found was Kafri *et al.* [26] who analyzed the differences in fitness between transcription and transcriptional-and-translation of a fluorescent protein driven by the pTDH3 promoter in yeast. This promoter is one of the strongest in yeast. Based on the RLTO model, we would predict this promoter to have a very high translation efficiency and therefore a large load ratio; however, the translation efficiency is much lower than one would predict based on a global analysis and likewise its load ratio is roughly unity, which based on the smaller than expected translation efficiency is broadly consistent with our expectations. A satisfactory test of this prediction will require larger-scale measurements that probe more representative genes.

8. Increased protein load analysis

Although the overall trend in the relation between the translation efficiency and message number in eukaryotic cells is captured by the RLTO model, a significant amount of scatter is observed around this optimal relation. One important consideration in a more realistic model are proteins whose fitness cost is greater than their metabolic load. (For instance, consider an ATPase with non-specific activity.) A second potentially interesting question is how message and protein length affect the optimal parameters. For instance, are a 50- and 500-amino-acid proteins expressed using a different regulatory program?

To explore the consequences of these added complexities, we can modify the metabolic load term in the growth rate equation (Eq. 4):

$$E\mu_m \to \lambda_p E\mu_m, \tag{A44}$$

which includes an additional parameter: the protein cost λ_p, which is 1 if the fitness cost is equal to the metabolic load and greater than one if the cost is higher. We will also treat the metabolic load per message λ as a gene specific parameter in this section only. The optimization can be repeated for this augmented model.

a. Derivation

To analyze the affect of increased protein load, we modify Eq. 4:

$$\ln \tfrac{k}{k_0} = -(\Lambda + \lambda_p E)\mu_m - \tfrac{1}{\ln 2}\gamma(\tfrac{\mu_m}{\ln 2}, \tfrac{\phi_p}{E \ln 2}), \tag{A45}$$

to contain the supplemental load factor λ_p which is unity if the only protein load is metabolic and $\lambda_p > 1$ if there is additions load (*e.g.* toxicity). The optimization conditions (Eqs. A27 and A28) become:

$$0 = -\tfrac{\lambda+\lambda_p\hat{\varepsilon}}{N_0} - \tfrac{1}{(\ln 2)^2}\gamma_{,1}(\tfrac{\hat{\mu}_m}{\ln 2}, \tfrac{n_p}{\hat{\varepsilon}\ln 2}), \tag{A46}$$

$$0 = -\tfrac{\lambda_p\hat{\mu}_m}{N_0} + \tfrac{n_p}{\hat{\varepsilon}^2(\ln 2)^2}\gamma_{,2}(\tfrac{\hat{\mu}_m}{\ln 2}, \tfrac{n_p}{\hat{\varepsilon}\ln 2}). \tag{A47}$$

Using the same algebraic approach as before, we can derive the same optimal overabundance and load equations:

$$\Lambda \ln 2 = -\partial_{\hat{\mu}_m}\gamma(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}), \tag{A48}$$

$$\hat{R} = \tfrac{1}{\Lambda \ln 2} p_\Gamma(\hat{\mu}_m | \tfrac{\hat{\mu}_m}{\ln 2}, \hat{o}\ln 2); \tag{A49}$$

however, the relation between the load and the translation efficiency now has an extra factor: λ_p:

$$R = \frac{\lambda_p \varepsilon \mu_m}{\lambda \mu_m} = \frac{\lambda_p \varepsilon}{\lambda}, \tag{A50}$$

representing the modified total load ratio.

b. Increased protein cost reduces the optimal translation efficiency.

The relation between the overabundance and message number is unchanged (Eq. 6). This result can be rationalized in the following way: The optimal overabundance is determined by the noise which is determined by message number only. This relation is unaffected by the added parameter λ_p. However, the optimal translation efficiency is affected:

$$\hat{\varepsilon} = \tfrac{\lambda}{\lambda_p}\hat{R}, \tag{A51}$$

where $\hat{R}$ is the optimal load ratio, defined by Eq. 10. The optimal curves are shown in Fig. 14.

How do these added considerations affect the RLTO predictions? First, we consider message and protein length. What are the optimal translation efficiencies for two proteins, one ten times the length of the other, at

Gene name	Message number: μ_m	Annotated function from Ecocyc	Essential (E)/ Nonessential (N) Ref. [45], [49], [47]
alsK	0.3	The alsK gene encodes a D-allose kinase. Its role in the degradation of D-allose is unclear; AlsK is not required for utilization of a D-allose carbon source; this effect may be due to the presence of other ambiguous sugar kinases within *E. coli* K-12.	E, N, N
bcsB	0.4	BcsB is encoded in a predicted operon together with *bcsA*, *bcsZ* and *bcsC*. In other organisms, these genes are involved in cellulose biosynthesis, a characteristic of the rdar (red, dry and rough) morphotype. However, the K-12 laboratory strain of *E. coli* does not show a rdar morphotype and does not produce cellulose.	E, N, N
entD	0.4	AcpS is the founding member of a 4'-phosphopantetheinyl (P-pant) transferase protein family that includes *E. coli* EntD, *E. coli* o195 protein, and *Bacillus subtilis* Sfp; family members share two conserved motifs but relatively low sequence identity overall.	E, N, N
yafF	0.4	No information about this protein was found by a literature search conducted on April 19, 2017.	E,-, N
yagG	0.6	*yagGH* is predicted to be a member of the XylR regulon; its products may mediate transport (YagG) and hydrolysis (YagH) of xylooligosaccharides; putative XylR and CRP binding sites are identified upstream of *yagGH*.	E,-, N
yceQ	0.2	No information about this protein was found by a literature search conducted on July 12, 2017.	E, E, N
ydiL	0.2	No information about this protein was found by a literature search conducted on April 7, 2017.	E, N, N
yhhQ	0.4	YhhQ is an inner membrane protein implicated in the uptake of queuosine (Q) precursors - 7-cyano-7-deazaguanine (preQ0) and 7-aminomethyl-7-deazaguanine (*preQ1*) - for Q salvage. Q-modified tRNA is absent in Δ*queD* and Δ*queD* Δ*yhhQ* strains grown in minimal media with glycerol; Q-modified tRNA is detected when a Δ*queD* strain is grown in minimal media plus 10 nM *preQ0* or *preQ1* but is absent when a Δ*queD* Δ*yhhQ* strain is grown under these conditions. *yhhQ* expressed from a plasmid restores the presence of Q-modified tRNA in a Δ*queD* Δ*yhhQ* strain.	E,-, N
yibJ	0.3	No information about this protein was found by a literature search conducted on July 9, 2018.	E, N, N
ymfK	0.4	YmfK is a component of the relic lambdoid prophage e14 and is likely the SOS-sensitive repressor. It is similar to the P34 gene of the *Shigella flexneri* bacteriophage SfV and belongs to the LexA group of SOS-response transcriptional repressors.	E, E, E

TABLE II. **Below-threshold essential genes identified in *E. coli*.** This table describes the message numbers and annotations for essential genes that we estimated to have expression below the threshold of one message per cell cycle. However, in the final column, we show classifications from three different studies. Only one of the identified genes, *ymfK*, was consistently defined as essential.

fixed protein number? In this case, we will assume that both the transcriptional cost (λ) as well as the translational cost (λ_p) increase tenfold. These increases cancel, resulting in the same optimal translation efficiency since it is only the relative cost of transcription to translation that is determinative of the translation efficiency.

Now consider a tenfold protein-specific increase in protein cost at fixed message cost and fixed protein number. The message number and translation efficiency would change by compensatory factors of 10:

$$\hat{\mu}_m \rightarrow 10 \cdot \hat{\mu}_m, \tag{A52}$$
$$\hat{\varepsilon}_m \rightarrow \tfrac{1}{10} \cdot \hat{\varepsilon}_m, \tag{A53}$$

to maintain the protein number.

Returning to our original motivation, we can understand how genes with a higher protein-to-message cost migrate downwards and rightwards off the optimal $\lambda_p = 1$ curve, predicting a cloud versus a narrow strip in proteome fraction measurements shown in Fig. 7. If the relative load Λ were directly measured, we would expect the predicted optimal translation efficiency curve for $\lambda_p = 1$ to lie at the top edge of the observed data cloud rather than the bisecting it. This bisection is the consequence of fitting an effective relative load parameter to the abundance data in the unaugmented RLTO model.

9. Analysis of translational limits.

A critical assumption in the RLTO model to this point has been that the optimal central dogma parameters are realizable in the cell; however, translation can be limited by a number of different mechanisms. The superior performance of the constant- over the optimal-translation-

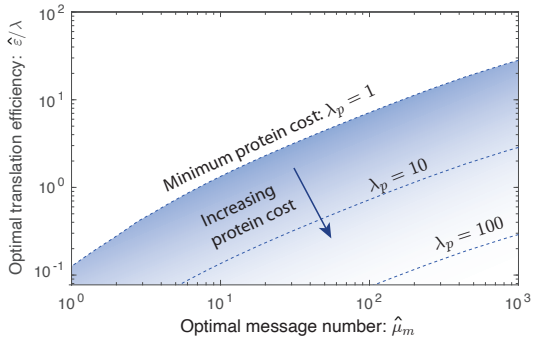

FIG. 14. **Increased protein cost decreases optimal translation efficiency.** A protein cost of $\lambda_p = 1$ corresponds to the metabolic cost of protein synthesis only, and is the minimum protein cost. For larger protein costs, the optimal translation efficiency is lower. As a result, the $\lambda_p = 1$ curve represents an upper bound of the optimal translation efficiency.

efficiency model in *E. coli* (Fig. 7D) suggests that this assumption may not be satisfied for bacteria. How do translation limits affect the model phenomenology?

When considering possible limits on translation, there are two natural mechanisms: (i) ribosome-number limit, where the number of ribosomes in the cell limits translation and (ii) a ribosome-per-message limit, where the number of ribosome per message is limiting. Assuming the ribosome-number-limit mechanism, the original unconstrained optimization problem can be recast as a constrained optimization problem where the protein cost λ_p is reinterpreted as a Lagrange multiplier to constrain the number of proteins translated (*e.g.* [50]). In spite of this reformulation, we would still predict the same functional form for the coupling between the optimal translation efficiency and message number. (*I.e.* it is still optimal to have a higher translation efficiency for highly-expressed genes even if the total number of proteins is fixed.) Therefore, the ribosome-number-limit mechanism cannot be the rationale for the constant translation efficiency observed in *E. coli*.

Assuming the ribosome-per-message-limit mechanism, we limit the translation efficiency to a restricted range of values. If the unconstrained optimum lies above this range, the optimum is at the maximum limiting value. If the unconstrained optima for all genes lie above the realizable range, the model predicts a translation efficiency uncoupled from the message number, as observed. These predictions are consistent with the observed central dogma regulatory program in *E. coli*. In added support of this hypothesis, Hausser *et al.* have argued that *E. coli* translates close to just such a ribosome-per-message limit as a consequence of the finite ribosome complex footprint on a message [14].

10. Parameter values for figures

The parameter values for the RLTO model used for each figure in the paper are shown in Tab. III. All calculations are independent of $\delta\tau$ and L_0.

11. Estimate of the message cost and metabolic load

We can estimate the message cost λ from the known total protein number for yeast and mammalian cells. (For *E. coli* this estimate is not possible since the protein cost in not determinative of the translation efficiency.)

The optimal translation efficiency is (Eq. A31):

$$\hat{\varepsilon} = \frac{\lambda}{\Lambda \ln 2} p_\Gamma(\hat{\mu}_m | \frac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2), \tag{A54}$$

and therefore the optimal protein number is:

$$\hat{\mu}_p = \hat{\mu}_m \hat{\varepsilon} = \frac{\lambda}{\Lambda \ln 2} \hat{\mu}_m p_\Gamma(\hat{\mu}_m | \frac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2). \tag{A55}$$

We define the normalization constant A:

$$A = \sum_i [\hat{\mu}_m \cdot \frac{1}{\Lambda \ln 2} p_\Gamma(\hat{\mu}_m | \frac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2)]_i, \tag{A56}$$

where the index i runs over all genes. Now, by summing Eq. A55, over all genes, we derive an expression for the total protein number N_p^{tot} in terms of the message cost λ and the normalization constant A:

$$N_p^{\text{tot}} = \lambda A. \tag{A57}$$

Solving for the protein cost results in the estimate:

$$\hat{\lambda} = \frac{N_p^{\text{tot}}}{A}. \tag{A58}$$

This message cost estimate $\hat{\lambda}$ can then be plugged into the metabolic load definition (Eq. A26) to estimate the multiplicity:

$$\hat{N}_0 \equiv L_0 + \hat{\lambda} N_m^{\text{tot}} + N_p^{\text{tot}} \tag{A59}$$

where we have ignored the non-protein and non-message contributions to the load ($L_0 = 0$).

a. *Detailed protocol*

We first estimate the message numbers, as described in Sec. C 2 c, from RNA-seq data. For each gene i, we set the optimal message number equal to the observed message number and then compute the optimal overabundance from the message number using Eq. A34. (Since the result is independent of the assumed Λ value, we set an arbitrary initial value of $\Lambda = 10^{-5}$.) We then use these single gene optimal message number and overabandances to compute A using Eq. A56. In Eqs. A58 and A59, we use the N_p^{tot} from Tab. IV. N_m^{tot} is computed by summing the estimated message numbers.

Figure number and panel:	Relative load: Λ (No units)	Protein cost: λ (molecules^{-1})	Protein threshold: n_p (molecules)	Message number: μ_m (molecules/cell cycle)	Translation efficiency: ε (No units)	Noise floor C_0 (No units)
3B	10^{-5}	0.1	10^2	Range	Range	0
3C	10^{-5}	Range	Range	Local optimum	Local optimum	0
4A	Range	NA	Range	Local optimum	Local optimum	0
4B	10^{-5}	NA	Range	Local optimum	30	0.1
5	Range	NA	Range	Local optimum	Local optimum	0
6A	Range	NA	Range	Local optimum	Local optimum	0
7B	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
7C	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
7D	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
8B	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
11A	10^{-5}	10	10^2	Range	Range	0
11B	10^{-5}	10	10^2	Range	Range	0
11C	10^{-5}	10	10^2	Range	Range	0
11D	10^{-5}	10	10^2	Range	Range	0
12A	Range	10^2	Range	Local optimum	Local optimum	0
12B	Range	10^2	Range	Local optimum	Local optimum	0
12C	Range	10^2	Range	Local optimum	Local optimum	0
12D	Range	10^2	Range	Local optimum	Local optimum	0
13A	Range	10^2	Range	Local optimum	Local optimum	0
13B	Range	10^2	Range	Local optimum	Local optimum	0
13C	Range	10^2	Range	Local optimum	Local optimum	0
13D	Range	10^2	Range	Local optimum	Local optimum	0
14	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
15	Range	10^2	Range	Local optimum	Local optimum	0

TABLE III. **RLTO model parameters by figure.** *Range* appears if a range of parameters is used. *NA* appears if the parameter value is irrelevant. *Local optimum* appears if the parameter values in optimized to maximize the growth rate. All calculations are independent of $\delta\tau$ and L_0.

b. *Estimate the message cost and metabolic load in yeast*

In yeast, the estimates are:

$$A = 4.8 \times 10^5, \quad \text{(A60)}$$
$$\hat{\lambda} = 1.0 \times 10^2, \quad \text{(A61)}$$
$$\hat{N}_0 = 6.2 \times 10^7, \quad \text{(A62)}$$
$$\hat{\Lambda} = 1.6 \times 10^{-6}, \quad \text{(A63)}$$

where the data sources are described in detail in Sec. C 1 b.

c. *Estimate the message cost and metabolic load in human cells*

In human cells, the estimates are:

$$A = 4.3 \times 10^6, \quad \text{(A64)}$$
$$\hat{\lambda} = 7.1 \times 10^2, \quad \text{(A65)}$$
$$\hat{N}_0 = 2.4 \times 10^9, \quad \text{(A66)}$$
$$\hat{\Lambda} = 2.9 \times 10^{-7}. \quad \text{(A67)}$$

where the data sources are described in detail in Sec. C 1 c.

12. Prediction of the proteome fraction

a. *RLTO: proteome fraction*

Starting from Eq. A55, clearly:

$$\hat{\mu}_p \propto \hat{\mu}_m \; p_\Gamma(\hat{\mu}_m | \tfrac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2), \quad \text{(A68)}$$

which can be used to predict the proteome fraction:

$$\hat{\Phi}_{pi} \equiv \tfrac{\hat{\mu}_{pi}}{\sum_j \hat{\mu}_{pj}}, \quad \text{(A69)}$$

where the second subscript is the gene index. To predict the proteome fraction, we computed the proportionality constant C:

$$C \equiv \textstyle\sum_i \left[\mu_m p_\Gamma(\mu_m | \tfrac{\mu_m}{\ln 2}, o \ln 2)|_{o=\hat{o}(\mu_m)}\right]_i, \quad \text{(A70)}$$

where the message numbers μ_{mi} for gene i are the experimentally observed message numbers, the implicit o_i values are predicted by the RLTO model (Eq. A33) for message number μ_{mi} and the sum index i runs over all genes. The predicted optimal proteome fraction is:

$$\hat{\Phi}_{pi} = C^{-1} \left[\mu_m p_\Gamma(\mu_m | \frac{\mu_m}{\ln 2}, o \ln 2)|_{o=\hat{o}(\mu_m)}\right]_i, \quad \text{(A71)}$$

which generates the predicted solid curves shown in Fig. 7ABC.

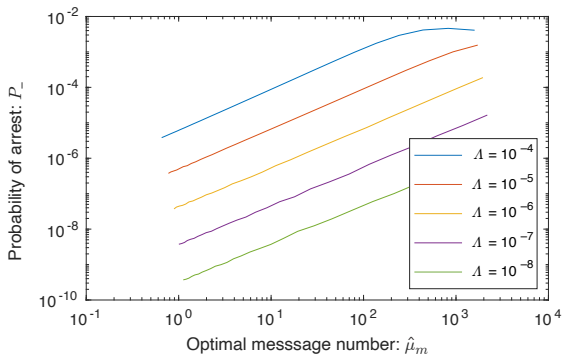

FIG. 15. **Cell cycle arrest probability.** Low protein number slows growth by arresting the cell cycle. The per-essential-gene arrest probabilities are shown as a function of message number μ_m. The arrest probability has an approximately linear dependence of the message number μ_m. The arrest probability is roughly proportion to the relative load Λ.

b. *Constant-translation-efficiency model: proteome fraction*

For the constant translation efficiency model, we define the normalation:

$$C' \equiv \sum_i \mu_{mi}, \tag{A72}$$

and the predicted proteome fraction is:

$$\Phi'_{pi} = C'^{-1} \mu_{mi}, \tag{A73}$$

which generates the predicted dotted curves shown in Fig. 7ABC.

c. *Sources of experimental data for proteome fraction analysis*

For *E. coli* data, the protein abundance data was generated by mass spec measurements and the message abundance data was from RNA-seq measurements [24]. For the yeast data, the protein abundance data is measured by mass spec and message abundances are determined by RNA-seq [28]. For the mammalian data, we used mouse data. The protein abundance data is measured by mass-spec and message abundances are determined by RNA-seq [29].

We estimated the message number μ_m as described in Sec. C 2 c. For the mouse data, the study provided message lifetimes, the cell cycle duration and abundances in molecules per cell [29]. For the *E. coli* and yeast, the total number of proteins, messages *etc*, cell cycle duration and message lifetimes for each organism and their sources are described in Tab. IV.

Appendix B: Analysis of alternative models

In this section, we investigate the phenomenology of three different single-cell growth rate functions to determine what model features result in overabundance. We consider a *threshold-like model* (the RLTO model), a *slow-growth model*, and a *symmetric model*. In each case, we will assume that the protein number is described by a gamma distribution:

$$N_p \sim \Gamma(\tfrac{\mu_m}{\ln 2}, \varepsilon \ln 2). \tag{B1}$$

We will assume the cell-cycle duration T is determined by this stochastic protein number N_p and then compute the population growth rate using Eq. A16 for a range of different message numbers μ_m. In each case, $\tau_0 = 1/N_0$, $N_0 = 10^5$, $\varepsilon = 30$, $n_p = \varepsilon \ln 2$. The mean expression level is $\mu_p = \mu_m \varepsilon$.

1. Threshold (RLTO) model

For the threshold-like (RLTO) model has cell cycle duration:

$$T = \tau_0 \begin{cases} \infty, & N_p < n_p \\ N_0 + N_p, & N_p > n_p \end{cases}, \tag{B2}$$

where protein expression below threshold n_p results in growth arrest.

2. Model 2: Slow-Growth model

In the slow-growth model, we imagine two processes: (i) checkpoint process X and (ii) other processes. The cell will divide after whichever process finishes last. Other processes will finish after time predicted by the metabolic load, identical to the threshold model defined above. However, we model checkpoint process X as the completion of a fixed amount of activity in an irreversible process. We will therefore assume it will take a time inversely proportional to the amount of enzyme X (N_p). The amount of activity is set by effective threshold n_p:

$$T = \tau_0 \max\{N_0 + N_p, \tfrac{2 n_p N_0}{N_p}\} \tag{B3}$$

such that n_p defines the level of protein required to make the growth rate half the metabolic limit.

Unlike the threshold model, cell growth slows but does not stop for $N_p < n_p$. This model will test whether are results are an artifact of the assumed-arrest based slow growth.

3. Model 3: Symmetric model

For the symmmetric model, we choose the model parameters such that the single-cell optimum was close to

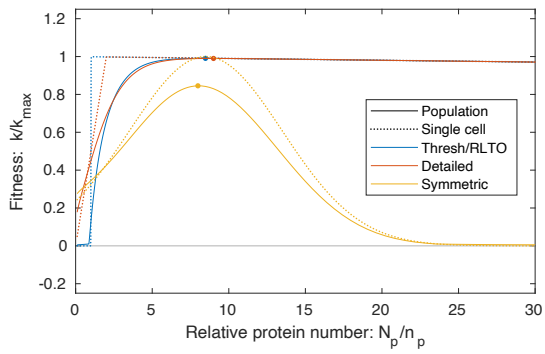

FIG. 16. **Exploring the mathematical mechanism of overabundance.** Single-cell and population growth rate are compared for three different models: threshold-like, slow-growth, and symmetric models. In the threshold-like model (RLTO), the growth rate goes to zero below threshold protein level n_p. In the slow-growth model, the growth rate transitions continuously to zero as the N_p is depleted below n_p. In both the threshold-like and slow-growth models, there is a small negative slope above the threshold corresponding to the metabolic load. In the symmetric model, the fitness cost is symmetric about the optimum. Both the threshold-like and slow-growth models are optimized at mean expression levels μ_p far exceeding the threshold level n_p. This is a consequence of the highly-asymmetric dependence of the fitness on protein number N_p. This leads to the phenomenon of protein overabundance. In contrast, the symmetric model is optimized in close proximity to its single-cell optimum.

the other models: $n_0 = 8.5$, $\sigma_n = 5$. The cell-cycle duration is

$$T = \tau_0 N_0 \exp\left(\frac{(N_p - n_0)^2}{2\sigma_n^2}\right) \tag{B4}$$

such that the noise-free growth rate will be Gaussian is N_p.

4. Conclusions from fitness-landscape analysis

The growth rates as a function of the mean expression level μ_p are shown in Fig. 16. The symmetric model has a population optimum in close proximity to its single-cell optimum, as we intuitively expect. However, both the threshold-like (RLTO) model and the slow-growth model have optima far above the threshold number n_p. We therefore conclude that it is fitness asymmetry rather than the threshold-like growth rate that is responsible for the overabundance phenomenon.

Why doesn't growth arrest of a sub-population lead to a stronger effect than the same sub-population growing slowly? In Ref. [23], we showed that the population doubling time $\overline{T}$ can be understood as the exponential mean of the stochastic cell-cycle duration:

$$\overline{T} \equiv f^{-1}[\mathbb{E}_T f(T)], \tag{B5}$$

where $\mathbb{E}_T$ is the expectation over the stochastic duration T and $f(t) \equiv \exp(-kt)$, where $k = \overline{T}^{-1} \ln 2$ is the population growth rate. Due to the functional form of $f(t)$, any long cell cycles are exponentially suppressed in their contribution to the exponential mean. Therefore, low-probability extremely-long-duration cell cycles only contribute to the growth rate by reducing the faction of growing cells.

Appendix C: Quantitation of central dogma parameters for one-message-rule

1. Selection of central dogma parameter estimates

The estimates for central dogma model parameters come from two types of data: (i) quantitative measurement of cellular-scale parameters for each organism (total number of messages in the cell, cell cycle duration, *etc*) and (ii) genome-wide studies quantitative of mRNA and protein abundance.

For the cellular-scale central dogma parameters, we relied heavily on an online compilation of biological numbers: BioNumbers [51]. This resource provides a collection of curated quantitative estimates for biological numbers, as well as their original source. In the interest of conciseness, we have cited only the original source in the Tab. IV, although we are extremely grateful and supportive of the creators of the BioNumbers website for helping us very efficiently identify consensus estimates for the parameters of the central dogma parameters.

For the selection of genome-wide studies on abundance, we used many of the same resources cited in BioNumbers as well as studies selected by a previous study of a quantitative analysis of the central dogma: Hausser *et al.* [14].

a. E. coli data

Message lifetimes: The message lifetimes (and median lifetime) were taken from a recent transcriptome-wide study by Chen *et al.* [41]. These investigators measured the lifetime in both rapid (LB) and slow growth (M9).

Noise: Taniguchi *et al.* have performed a beautiful simultaneous study of the proteome and transcriptome with single-molecule sensitivity [13]. Although we use the noise analysis data from this study for our supplemental analysis of *E. coli* noise, it is not the source for our *E. coli* transcriptome data due to the extremely slow growth of the cells in this study (150 minute doubling time), which is not consistent with the growth conditions for the other sources of data.

mRNA abundance: Instead, we used data from the more recent Bartholomaus *et al.* study [52], which char-

acterizes the transcriptome in both rapid (LB) and slow growth (M9).

Total cellular message number. This study was chosen since it was the source of the BioNumbers estimates of cellular message number in *E. coli* (BNID 112795 [51]).

Doubling time: The source of the doubling times for rapid (LB) and slow (M9) growth of *E. coli* comes from Bernstein [40].

Essential gene classification. The classification of essential genes in yeast comes from the construction of the Keio knockout collection from Baba *et al.* [45].

Protein number. The total protein number in *E. coli* came from Milo's recent review of this subject [53].

b. Yeast data

Message lifetimes: The message lifetimes (and median lifetime) were taken from Chia *et al.* [54].

Noise: The noise data was taken from the Newman *et al.* study, which used flow cytometry of a library of fluorescent fusions to characterize protein abundance with single-cell resolution [12].

mRNA abundance: The transcriptome data comes from the very recent Blevins *et al.* study [55].

Total cellular message number. There are a wide-range of estimates for the total cellular message number in yeast: 1.5×10^4 [56] (BNID 104312 [51]), 1.2×10^4 [57] (BNID 102988 [51]), 6.0×10^4 [58] (BNID 103023 [51]), 2.6×10^4 [59] (BNID 106763 [51]) and 3.0×10^4 [60]. We used the compromise value of 2.9×10^4.

Doubling time: The doubling time was taken from [61].

Protein number. The total protein number in yeast comes from Futcher *et al.* [62].

Essential gene classification. The classification of essential genes in yeast comes from van Leeuwen *et al.* [63].

Proteome abundance data: The proteome abundance data came from two sources: flow cytometry of fluorescent fusions from Newman *et al.* [12] as well as mass-spec data from de Godoy *et al.* [64].

c. Human data

Message lifetimes: The message lifetimes (and median lifetime) were taken from Yang *et al.* [65] who reported a median half life of 10 h which corresponds to a lifetime of 14 h.

mRNA abundance: The transcriptome data comes from the data compiled by the Human Protein Atlas [66], which we averaged over tissue types.

Total cellular message number. The total cellular message number in human comes from Velculescu *et al.* [67] (BNID 104330 [51]).

Doubling time: The doubling time was taken from [68].

Protein number. The total protein number in human came from Milo's recent review of this subject [53].

Essential gene classification. The classification of essential genes in human comes from Wang *et al.* [69].

2. Quantitative estimates of central dogma parameters

a. Estimating the cellular message number: $\mu_{m/c}$

For each model organism (and condition), we found a consensus estimate from the literature for the total number of mRNA messages per cell $N^{\text{tot}}_{m/c}$. This number and its source are provided in Tab. IV. To estimate the number of messages corresponding to gene i, we re-scaled the un-normalized abundance level r_i:

$$N_{m/c,i} = N^{\text{tot}}_{m/c} \frac{r_i}{\sum_j r_j}, \tag{C1}$$

where the sum over gene index j runs over all genes.

b. Estimating the transcription rate: β_m

To estimate the transcription rate for gene i, we start from the estimated cellular message number $N_{m/c,i}$ and use the telegraph model prediction for the cellular message number:

$$N_{m/c,i} = \beta_{m,i}/\gamma_{m,i}, \tag{C2}$$

where $\gamma_{m,i}$ is the message decay rate. Since gene-to-gene variation in message number is dominated by the transcription rate (*e.g* [41]), we estimate the decay rate as the inverse gene-median message life time:

$$\gamma_{m,i} = \tau_m^{-1}, \tag{C3}$$

for which a consensus value was found from the literature. This number and its source are provided in Tab. IV. We then estimate the gene-specific transcription rate:

$$\beta_{m,i} = N_{m/c,i}/\tau_m. \tag{C4}$$

c. Estimating the message number: μ_m

To estimate the message number of gene i, we use the predicted value from the telegraph model:

$$N_{m,i} = T\beta_{m,i} = \frac{T}{\tau_m} N_{m/c,i}, \tag{C5}$$

where T is the doubling time and $N_{m/c,i}$ is the cellular message number (Eq. C1).

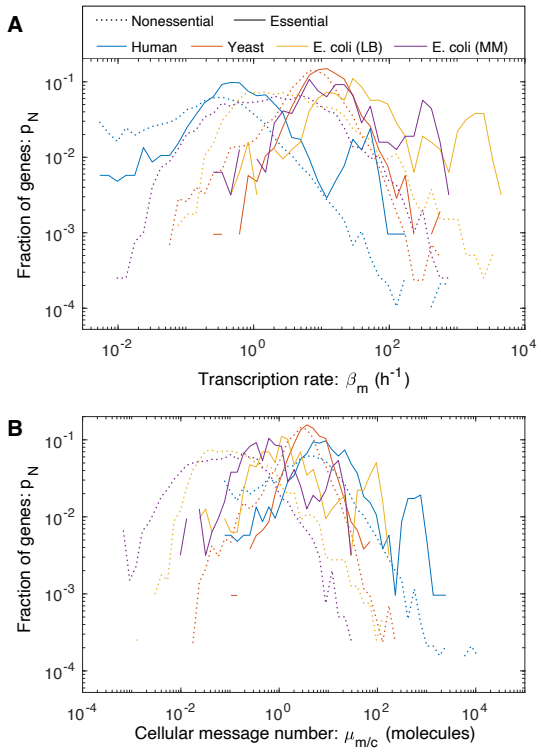

FIG. 17. **Transcription in three model organisms.** We characterized different gene transcriptional statistics in three model organisms. In *E. coli*, two growth conditions were analyzed: Rapid growth in LB and Slow growth in minimal media. **Panel A: The distribution of gene transcription rate.** The typical transcription rate varies by two orders-of-magnitude between organisms. **Panel B: The distribution of gene cellular message number.** There is also a two-order-of-magnitude variation between typical cellular message numbers. No consistent lower threshold is observed for either statistic.

d. *Histograms*

We generated histograms for each of the three transcriptional statistics: transcription rate β_m, cellular message number $\mu_{m/c}$, and message number μ_m. The histograms for transcription rate and cellular message number do not show a consistent lower limit (as predicted) and are shown in Fig. 17; however, the histogram for message number does show a consistent lower bound for the three model organisms and is shown in Fig. 5B.

Appendix D: Supplemental analysis of Noise-Protein-Abundance Relation in Yeast

1. Estimating *protein number* (μ_p) for the noise analysis

The protein abundance data for yeast grown in YEPD media and measured with flow cytometry fluorescence [12] were given in arbitrary units (AU). In order to convert from AU to protein number, the fluorescence values were rescaled by comparing with mass-spectrometry protein abundance data for yeast grown in YNB media [64]. Since the protein abundance from mass-spectrometry was given in terms of Intensity, the Intensity values were first rescaled by the total number of proteins in yeast, 5×10^7. (See Sec. C 1 b.) The mass-spectrometry protein data was thresholded at 10 proteins, based on the assumption that the noise of the data for 10 and fewer proteins makes the data unreliable. Next, the log of the fluorescence protein abundance in AU as a function of the log of thresholded mass-spectrometry protein abundance was fit as a linear function with an assumed slope of 1 to find the offset, 3.9, (Fig. 18) which corresponds to a multiplicative scaling factor. We then used that offset value to rescale the fluorescence data from AU to protein number. We also compared to yeast grown in SD media [12] and found a similar offset result.

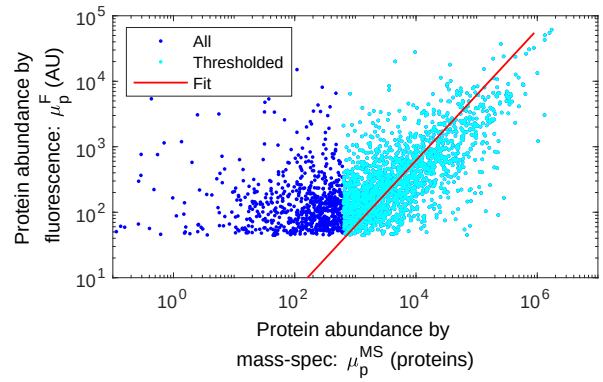

FIG. 18. **Fit to rescale fluorescence intensity to protein number.** Protein abundance from flow cytometry fluorescence [12] as a function of mass-spectrometry scaled abundance [64]. The mass-spectrometry data was thresholded at 10 proteins, and then a linear fit was performed to find the multiplicative offset of 3.9, which was used to convert protein fluorescence AU to number.

2. Empirical models for yeast gene expression

To generate the empirical model for protein number as a function of message number, we used protein abundance data from Newman *et al.* [12], re-scaled to estimate protein number (Sec. D 1) and transcriptome data from Lahtvee *et al.* [70], re-scaled to estimate message number (Sec. C 2 c).

a. *The meaning of the error estimates*

Before providing a detailed error analysis, it is important to place our error estimates in perspective. The error that we will be estimating is the statistical error

Model organism	Growth condition	Doubling time: T	Message lifetime: $\tau_m = \gamma_m^{-1}$	Message recycling ratio: $m = T/\tau_m$	Total number of messages /cell: $N_{m/c}^{\mathrm{tot}}$	Total number of messages /cell-cycle: N_m^{tot}	Total number of proteins: N_p^{tot}	Average translation efficiency: ε	Average translation rate: β_p (h^{-1})
Escherichia coli (*E. coli*)	LB	30 min [40]	2.5 min [41]	12	7.8×10^3 [52]	9.4×10^4	3×10^6 [53]	22	530
	M9	90 min [40]	2.5 min [41]	36	2.4×10^3 [52]	8.6×10^4	3×10^6 [53]	24	580
Sacchromyces cerevisiae (Yeast–haploid)	YEPD	90 min [61]	22 min [54]	4	2.9×10^4 [59]	1×10^5	5×10^7 [62]	4×10^2	410
Mus musculus (Mammalian-mouse)	Tissue	27.5 h [29]	15 h [29]	1.8	1.7×10^5 [29]	3×10^5 [29]	3×10^9 [29]	1×10^4	660
Homo sapiens (Human)	Tissue	24 h [68]	14 h [65]	1.7	3.6×10^5 [61]	5×10^5	2×10^9 [53]	4×10^3	120

TABLE IV. **Central dogma parameters for three model organisms with detailed references.** Columns three through seven hold representative values for measured central-dogma parameters for the model organisms described in the paper. Each value is followed by a reference for its source.

associated with the finite sample size; however, *this is not the dominant source of error*. A far more important consideration are systematic problems with our analysis and the underlying experiments. For instance, since we do not have a detailed model for the error of the experiments analyzed, there are multiple distinct analyses (*i.e.* assumptions about the error model) that could be implemented for the data fitting, each giving slightly different model parameters. These model to model differences still give rise to predictions consistent with our qualitative conclusions; however, they are likely larger than the statistical uncertainty we compute (while assuming a particular model).

b. *Empirical model for protein number*

We initially fit the empirical model for protein number,

$$\mu_p = C_0 \mu_m^{\alpha_0}, \tag{D1}$$

to the data using a standard least-squares approach; however, the algorithm led to a very poor fit since it does not account for uncertainty in both independent and dependent variables. We therefore used an alternative approach [71], which assumes comparable error in both variables. The model parameters are:

$$\begin{aligned} \alpha_0 &= 2.1 \pm 0.04, & \text{(D2)} \\ C_0 &= 8.0 \pm 1.0, & \text{(D3)} \end{aligned}$$

where the uncertainties are the estimated standard errors.

c. *Empirical model for message number*

For the prediction of the coefficient of variation, it is useful to invert Eq. D1 to generate a model for message number as a function of protein number:

$$\begin{aligned} \mu_m &= C_0^{-1/\alpha_0} \mu_p^{1/\alpha 0}, & \text{(D4)} \\ &= C_1 \mu_p^{\alpha_1}, & \text{(D5)} \end{aligned}$$

where the last line defines two new parameters: a coefficient C_1 and an exponent α_1. The resulting parameters and uncertainties are:

$$\begin{aligned} \alpha_1 &\equiv 1/\alpha_0, & \text{(D6)} \\ &= 0.48 \pm 0.01, & \text{(D7)} \\ C_1 &\equiv C_0^{-1/\alpha_0}, & \text{(D8)} \\ &= 0.37 \pm 0.02, & \text{(D9)} \end{aligned}$$

where the uncertainties are the estimated standard errors.

d. *Empirical model for translation efficiency*

To generate an empirical model for translation efficiency, we started from the empirical model for protein number (Eq. D1), and then use Eq. 1 to relate protein number, message number, and translation efficiency:

$$\begin{aligned} \varepsilon &= \frac{\mu_p}{\mu_m}, & \text{(D10)} \\ &= C_0 \mu_m^{\alpha_0 - 1}, & \text{(D11)} \\ &= C_2 \mu_m^{\alpha_2}, & \text{(D12)} \end{aligned}$$

where the last line defines two new parameters: a coefficient C_2 and an exponent α_2. The resulting parameters

and uncertainties are:

$$\alpha_2 = \alpha_0 - 1, \quad \text{(D13)}$$
$$= 1.07 \pm 0.04, \quad \text{(D14)}$$
$$C_2 = C_0, \quad \text{(D15)}$$
$$= 8.0 \pm 1.0, \quad \text{(D16)}$$

where the uncertainties are the estimated standard errors.

e. *Empirical model for the coefficient of variation*

To generate an empirical model for the coefficient of variation, we started from the empirical model for message number (Eq. D5), and then substitute this into the statistical model prediction for CV_p^2 (Eq. 2):

$$\mathrm{CV}_p^2 = \frac{\log 2}{\mu_m}, \quad \text{(D17)}$$
$$= C_0^{1/\alpha_0} \log 2 \cdot \mu_p^{-1/\alpha_0}, \quad \text{(D18)}$$
$$= C_3 \mu_p^{\alpha_3}, \quad \text{(D19)}$$

where the last line defines two new parameters: a coefficient C_3 and an exponent α_3. The resulting parameters and uncertainties are:

$$\alpha_3 \equiv -1/\alpha_0, \quad \text{(D20)}$$
$$= -0.48 \pm 0.01, \quad \text{(D21)}$$
$$C_3 \equiv C_0^{1/\alpha_0} \log 2, \quad \text{(D22)}$$
$$= 1.9 \pm 0.1, \quad \text{(D23)}$$

where the uncertainties are the estimated standard errors.

3. Supplemental analysis of gene expression noise

The quantitative model for gene expression noise includes multiple contributions:

$$\mathrm{CV}_p^2 \approx \frac{1}{\mu_p} + \frac{\log 2}{\mu_m} + c_0, \quad \text{(D24)}$$

where the first term can be understood to represent the Poisson noise from translation, the second term the Poisson noise from transcription, and the last term, c_0, is called the *noise floor* and is believed to be caused by the cell-to-cell variation in metabolites, ribosomes, and polymerases *etc* [72, 73].

a. *Inclusion of noise floor in the yeast analysis*

In the main text of the paper, we have ignored the role of the noise floor in the analysis of noise in yeast. Unlike *E. coli*, where the noise floor is high ($\mathrm{CV}_p^2 = 0.1$) and is determinative of the noise associated with almost all essential genes [13, 72, 73], in yeast the noise floor is much lower ($\mathrm{CV}_p^2 = 0.01$) and therefore affects only genes with the highest expression.

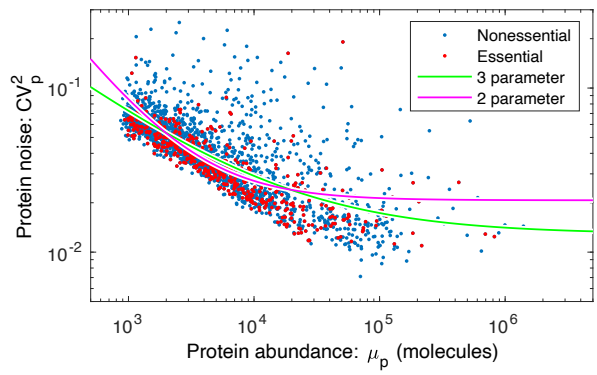

FIG. 19. **Yeast noise fit against canonical noise model, with a noise floor.** Yeast noise data fit with the 2- (null hypothesis with μ_p^{-1} dependence) and 3- parameter (μ_p^a) models. The two-parameter model corresponds to the canonical noise model (Eq. 13) and fails to quantitatively fit the data.

In this section, we will consider models that include the noise floor, since its presence can make the noise scaling more difficult to interpret. To determine if the scaling of the noise is consistent with the canonical assumption that the noise is proportional to μ_p^{-1} for low expression, we will consider two competing empirical models for the noise (Fig. 19). In the null hypothesis, we will consider a model:

$$\eta_0(\mu_p; b, c) = \frac{b}{\mu_p} + c, \quad \text{(D25)}$$

and an alternative hypothesis with an extra exponent parameter a:

$$\eta_1(\mu_p; a, b, c) = \frac{b}{\mu_p^a} + c. \quad \text{(D26)}$$

We will assume that CV_p^2 is normally distributed about η with unknown variance σ_η^2.

In this context, a maximum likelihood analysis is equivalent to least-squares analysis. Let the sum of the squares be defined:

$$S_I(\boldsymbol{\theta}) \equiv \sum_i [\mathrm{CV}_{p,i}^2 - \eta_I(\mu_{p,i}; \boldsymbol{\theta})]^2 \quad \text{(D27)}$$

for model I where $\boldsymbol{\theta}$ represents the parameter vector. The maximum likelihood parameters are

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} S_I(\boldsymbol{\theta}), \quad \text{(D28)}$$

with residual norm:

$$\hat{S}_I = S_I(\hat{\boldsymbol{\theta}}). \quad \text{(D29)}$$

To test the null hypothesis, we will use the canonical likelihood ratio test with the test statistic:

$$\Lambda \equiv 2 \log \frac{q_1}{q_0}, \quad \text{(D30)}$$

where q_0 and q_1 are the likelihoods of the null and alternative hypotheses, respectively. Wilks' theorem states that Λ has a chi-squared distribution of dimension equal to the difference of the dimension of the alternative and null hypotheses ($3-2=1$).

b. Hypothesis test I

In our first analysis, we will estimate the variance directly. We computed the mean-squared difference for successive CV_p^2 values, sorted by mean protein number μ_p. The variance estimator is

$$\hat{\sigma}_\eta^2 = \tfrac{1}{2}\left\langle (\mathrm{CV}_{p,i}^2 - \mathrm{CV}_{p,i+1}^2)^2 \right\rangle_i = 6.3\times 10^{-4}, \tag{D31}$$

where the brackets represent a standard empirical average over gene i for the μ_p-ordered gene CV_p^2 values. The test statistic can now be expressed in terms of the residual norms:

$$\begin{aligned}\Lambda &= (\hat{S}_1 - \hat{S}_2)/\hat{\sigma}_\eta^2, &\text{(D32)}\\ &= 3.3\times 10^4, &\text{(D33)}\end{aligned}$$

which corresponds to a p-value far below machine precision. We can therefore reject the null hypothesis.

c. Hypothesis test II

In a more conservative approach, we can use maximum likelihood estimation to estimate the variance of each model independently as a model parameter. In this case, the test statistic can again be expressed in terms of the residual norms:

$$\begin{aligned}\Lambda &= N\log\tfrac{\hat{S}_1}{\hat{S}_2}, &\text{(D34)}\\ &= 1.6\times 10^2, &\text{(D35)}\end{aligned}$$

where N is the number of data points. (Details of derivation are in Sec. D 3 e.) In this case, the p-value can be computed assuming the Wilks' theorem (*i.e.* the chi-squared test):

$$p = 6\times 10^{-36}, \tag{D36}$$

again, strongly rejecting the null hypothesis.

d. Maximum likelihood estimates of the parameters

In the alternative hypothesis, the maximum likelihood estimate (MLE) of the empirical noise model (Eq. D26) parameters are (Fig. 19):

$$\begin{aligned}a &= 0.57\pm 0.02, &\text{(D37)}\\ b &= 3.0\pm 0.5, &\text{(D38)}\\ c &= 0.013\pm 0.001, &\text{(D39)}\end{aligned}$$

where the parameter uncertainty has been estimated using the Fisher Information in the usual way using the MLE estimate of the variance.

e. Details: Statistical details MLE estimate of the variance

The minus-log-likelihood for the normal model I is:

$$h_I(\hat{\boldsymbol{\theta}}, \sigma^2) = \tfrac{N}{2}\log 2\pi\sigma^2 + \tfrac{1}{2\sigma^2}\hat{S}_I, \tag{D40}$$

where $\hat{S}_I$ is the least-square residual. We then minimize h_I with respect to the variance σ^2:

$$\partial_{\sigma^2} h|_{\hat{\sigma}^2} = 0, \tag{D41}$$

to solve for the MLE $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \tfrac{1}{N}\hat{S}_I. \tag{D42}$$

Next we evaluate h at the variance estimator:

$$h_I(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) = \tfrac{N}{2}\left[\log 2\pi\tfrac{\hat{S}_I}{N} + 1\right]. \tag{D43}$$

The test statistics can be written in terms of the h's:

$$\begin{aligned}\Lambda &= 2h_0(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) - 2h_1(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2), &\text{(D44)}\\ &= N\log\tfrac{\hat{S}_0}{\hat{S}_1}, &\text{(D45)}\end{aligned}$$

which can be evaluated directly in terms of the residual norms for the null and alternative hypotheses.

Appendix E: Comments on the Hausser *et al.* analysis

Hausser *et al.* have previously performed a more limited analysis of the trade-off between metabolic load and gene-expression noise [14]. In this section, we will provide some more context into the differences between the two approaches.

The Hausser model assumes the the growth rate has the form:

$$k = k_0 - \tfrac{1}{2}|k''|(N_p - \mu_p)^2 - k_0\Lambda\mu_m, \tag{E1}$$

where k is the growth rate and we have rewritten the form of the fitness to better match our own definitions. Here k'' is the second derivative of the growth rate at the optimal protein number μ_p and N_p is the stochastic protein number. If we take the expectation with respect to the protein number, we get:

$$k = k_0 - \tfrac{1}{2}|k''|\sigma_p^2 - k_0\Lambda\mu_m, \tag{E2}$$

and substitute the noise model for the variance of the protein number gives:

$$k = k_0 - \tfrac{1}{2}|k''|\mu_p^2(\tfrac{\ln 2}{\mu_m} + C_0) - k_0\Lambda\mu_m, \tag{E3}$$

where C_0 is the noise floor and we have assumed the mean protein number is optimal (μ_p). If we maximize the growth rate with respect to μ_m, we get the following condition on the optimal message number:

$$\hat{\mu}_m^2 = \tfrac{1}{2}\tfrac{|k''|}{k_0}\mu_p^2\tfrac{\ln 2}{\Lambda}, \tag{E4}$$

which depends on the unknown curvature k''. To make global predictions about how transcription and translation are related, some added assumptions are necessary to describe how k'' scales with protein abundance.

To illustrate how this expression does not make explicit global predictions, let's consider a number of plausible possibilities. First we will assume that k'' is independent of μ_p and on average all proteins are equally sensitive to changes in protein number. In this case, we find:

$$\mu_p \propto \hat{\mu}_m \sqrt{\Lambda}, \tag{E5}$$
$$\hat{\varepsilon} \propto \sqrt{\Lambda}, \tag{E6}$$

implying a constant translation efficiency which is inversely proportional to the square root of the relative load.

Alternatively, we can assume that $k'' \propto \mu_p^{-2}$ and, on average, the cell is equally sensitive to changes in the relative number of proteins (*i.e.* $\Delta p/\mu_p$), regardless of expression level. In this case,

$$\hat{\mu}_m \propto 1/\sqrt{\Lambda}, \tag{E7}$$
$$\hat{\varepsilon} \propto \mu_p \sqrt{\Lambda}, \tag{E8}$$

implying a constant message number, irrespective of expression level, and a translation efficiency that is proportional to expression level.

Finally, we will assume that $k'' \propto \mu_p^{-1}$, which is the intermediate case. Here:

$$\mu_p \propto \hat{\mu}_m^2 \Lambda, \tag{E9}$$
$$\hat{\varepsilon} \propto \hat{\mu}_m \Lambda, \tag{E10}$$

implying that translation efficiency should increase with message number, analogous to our prediction. It appears that Hausser *et al.* implicitly also favor this model, since they define their sensitivity parameter to include a power of protein number μ_p. They justify this assumption by arguing that since $\sigma_p^2 \propto \mu_p$, it makes sense to define the *sensitivity to noise* to include a factor of μ_p [14]. At best, this is somewhat fuzzy logic since, as we demonstrate in the paper, Eq. E10 implies that the protein variance does not scale $\sigma_p^2 \propto \mu_p$!

The authors also propose a lower limit on the translation-transcription ratio; however, their limit is dependent on the noise floor, which only affects genes with the highest transcription rates in eukaryotic cells. The implementation of a more appropriate estimate of the noise, relevant for the vast majority of genes, does not lead to the same limit.