When synthetic biology fails: a modular framework for modelling genetic stability in engineered cell populations

Duncan Ingram and Guy-Bart Stan, Imperial College London

28th November 2022

Abstract

Predicting the dynamics of mutation spread in engineered cell populations is a sought-after goal in synthetic biology. Until now, models that capture these processes have been lacking, either by failing to account for the diversity of mutation types, or by failing to link the growth rate of a cell to the consumption of shared cellular resources by synthetic constructs. In this study we address these shortcomings by building a novel mutation-aware modelling framework of cell growth in a turbidostat. Our framework allows users to input essential design elements of their synthetic constructs so as to predict the time evolution of different mutation phenotypes and protein production dynamics. Its structure allows quick mutation-based analysis of any construct design, from single-gene constructs to multi-gene devices with regulatory elements. We show how our framework can generate new insights into industrial applications, such as how the design of synthetic constructs impacts long-term protein yield and genetic shelf-life. Our framework also uncovers new mutation-driven design paradigms for synthetic gene regulatory networks, such as how mutations can temporarily increase the bistability of toggle switches, or how repressilators can be resistant to single points of failure.

1 Introduction

It has long been known that cells engineered with synthetic genes eventually lose their intended function due to the onset of DNA mutations. Until recently, however, tools to understand, predict and control these mechanisms have been lacking. Acquiring accurate insight into the dynamics of mutation spread in a cell population would enhance many aspects of synthetic biology's Design-Build-Test-Learn paradigm, such as boosting the protein yield of cells in bioreactors [1],

controlling biosafety when deploying cells into the environment [2], and forcing synthetic devices to evolve in predictable ways [3]. Having predictive control over these processes is, therefore, a highly sought-after goal.

In recent years, experimental work has been devoted to the development of techniques that constrain genetic evolution, from coupling a synthetic construct's function to an essential gene [4], to expressing synthetic genes using orthogonal pathways in order to minimise cell stress [5]. Despite this, little progress has been made on *predicting* how mutations arise and propagate in a population of engineered cells. Current discussions in synthetic biology are increasingly focusing on the importance of this, with a recent insight by Castle et al. [3] exploring the utility of building 'evolutionary landscapes' that would allow users to predict how synthetic devices evolve over time. Some groups have implemented tools to study isolated aspects of these landscapes, such as Pinheiro et al.'s [6] model that predicts evolutionary trajectories of cell growth rate and antibiotic resistance levels. Despite this, no general computational frameworks of mutation spread in engineered cell populations have emerged.

For a model to successfully describe mutation spread in a cell population, it needs to incorporate two essential components: (i) the genetic stability of the synthetic DNA, such as its associated 'mutation probability', and (ii) the selection pressures that act on cells, such as how fast mutant varieties grow relative to non-mutant. Genetic stability depends on many factors, from exogenous agents like UV radiation [7] to errors in DNA replication [8]. A number of computational tools attempt to capture these phenomena [9–11], however they do not provide insight into the functional effect of mutations in the context of a growing cell – a prerequisite to capturing population-wide selection pressures. Using a different approach, the combined effect of genetic stability and selection pressure has been captured by more recent models that consider how disabling a genetic construct's expression impacts cell growth rate [12, 13]. While this provides another layer of understanding, these models are ultimately constrained by the fact that they only describe single types of mutation, and account for changes to cell growth through a single fixed parameter. As a result, they are unable to capture the myriad of mutation dynamics seen in typical experimental data that track loss of construct function over time [14]. More accurate models, therefore, would need to capture mutation heterogeneity and describe cell growth with more granularity. These phenomena are even more important to account for in present-day synthetic biology where experimental designs have surged in complexity, from multi-gene regulatory networks [15] to co-cultures implementing metabolic division of labour [16].

The concept of mutation heterogeneity has been applied at different physiological scales, from describing phenotypic variety in cancer cells [17] to quantifying epigenetic variation [18]. When engineering cells with synthetic constructs, however, it is more relevant to consider where the mutation occurs ('location') and how debilitating it is ('severity'). Such concepts are yet to be captured in existing models of engineered cell populations. When accounting for cell growth, it is important to recognise that expressing additional proteins reduces the availability of shared cellular resources, such as nucleotide triphosphates and ribosomes. High expression levels can therefore stress the cell and lead to reduced growth rate [19, 20], a phenomenon that has since motivated the development of 'host-aware' [21] cell models that consider how resources are shared between host and synthetic constructs [22–25]. By combining these ideas, it can be argued that a modelling framework for mutation spread not only needs to consider genetic stability and selection pressure, but in doing so should also account for mutation heterogeneity and the host-aware effects of synthetic gene expression. No existing models achieve these aspects in unison, motivating the creation of a novel modelling framework that is (i) mutation-aware, (ii) host-aware, and (iii) accounts for varied gene construct designs inline with present-day synthetic biology. The mutation-aware computational framework we present here provides new insight into industrial applications, such as how the DNA design of synthetic constructs impacts long-term protein yield and genetic shelf-life. It also provides insight into the mutation-driven dynamics of classic synthetic constructs such as the toggle switch and the repressilator, and suggests how their designs can be optimised to enhance genetic stability.

2 Creating a modular mutation-aware framework

2.1 Combining host-aware cell modelling with mutation state transitions

The accumulation of cells with mutated synthetic DNA can be modelled by considering transitions between different mutation phenotypes over time. A turbidostat is chosen as the growth setting such that the number of cells is kept at a constant value, N, by diluting with fresh media. In its simplest form, the model encompasses two states: engineered cells (E-cells) of quantity E, which have fully-functioning synthetic DNA, and mutant cells (M-cells) of quantity M, whose synthetic DNA has been mutated and rendered inactive (Figure 1a). These cells grow at rates

λ_E and λ_M respectively, and all cells belonging to a particular state are assumed to be identical. During the course of a cell cycle, a mutation that fully inactivates the synthetic DNA can occur with probability z, leading to the production of one E-cell and one M-cell upon division. By considering the growth rate of each cell type, the division rate of an E-cell into one E-cell and one M-cell is $\lambda_E \cdot z$, while the division rate into two E-cells is $\lambda_E \cdot (1-z)$. Mutations are assumed to be irreversible such that M-cell division always produces two M-cells at a rate of λ_M.

When the optical density (OD) in a turbidostat surpasses a target value, cells are typically diluted by a fixed volume of media which causes the OD to drop. This creates a periodic 'rise and fall' of cell number around a fixed value over time. For simplicity, we approximate this process by assuming that dilution occurs instantaneously in response to cell growth, keeping the OD at the target value, and hence keeping the number of cells at a fixed value, N. This requires a dilution function 'dil', which when $E + M > N$, removes cells in proportion to their abundance above N:

$$\text{dil} = \begin{cases} E + M - N, & \text{if } E + M > N \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Based on these considerations, an Ordinary Differential Equation (ODE) model capturing the dynamics of each cell type can be formulated as follows:

$$\dot{E} = E \cdot \lambda_E \cdot (1 - z) - E \cdot \text{dil}, \tag{2}$$

$$\dot{M} = E \cdot \lambda_E \cdot z + M \cdot \lambda_M - M \cdot \text{dil}. \tag{3}$$

In order to simulate realistic cell behaviours, parameter values for growth rate should consider the impact of synthetic gene expression on host cell growth. We capture these effects by integrating the state transition equations (Equations (2)–(3)) with the host-aware cell model from Weisse et al. [22], chosen due to its comparative simplicity and ease of manipulation compared to other host-aware models [23–25]. In its base form, it uses a set of ODEs that describes how essential resources are distributed in an engineered *E. coli* cell (SI Equations (2a)–(2m)). Growth rate at any time point is then calculated from the current combined rate of producing all proteins in the cell (SI Equation 4). This dynamically changes based on how many resources are diverted from growth-supporting processes (such as producing ribosomes) to synthetic gene expression.

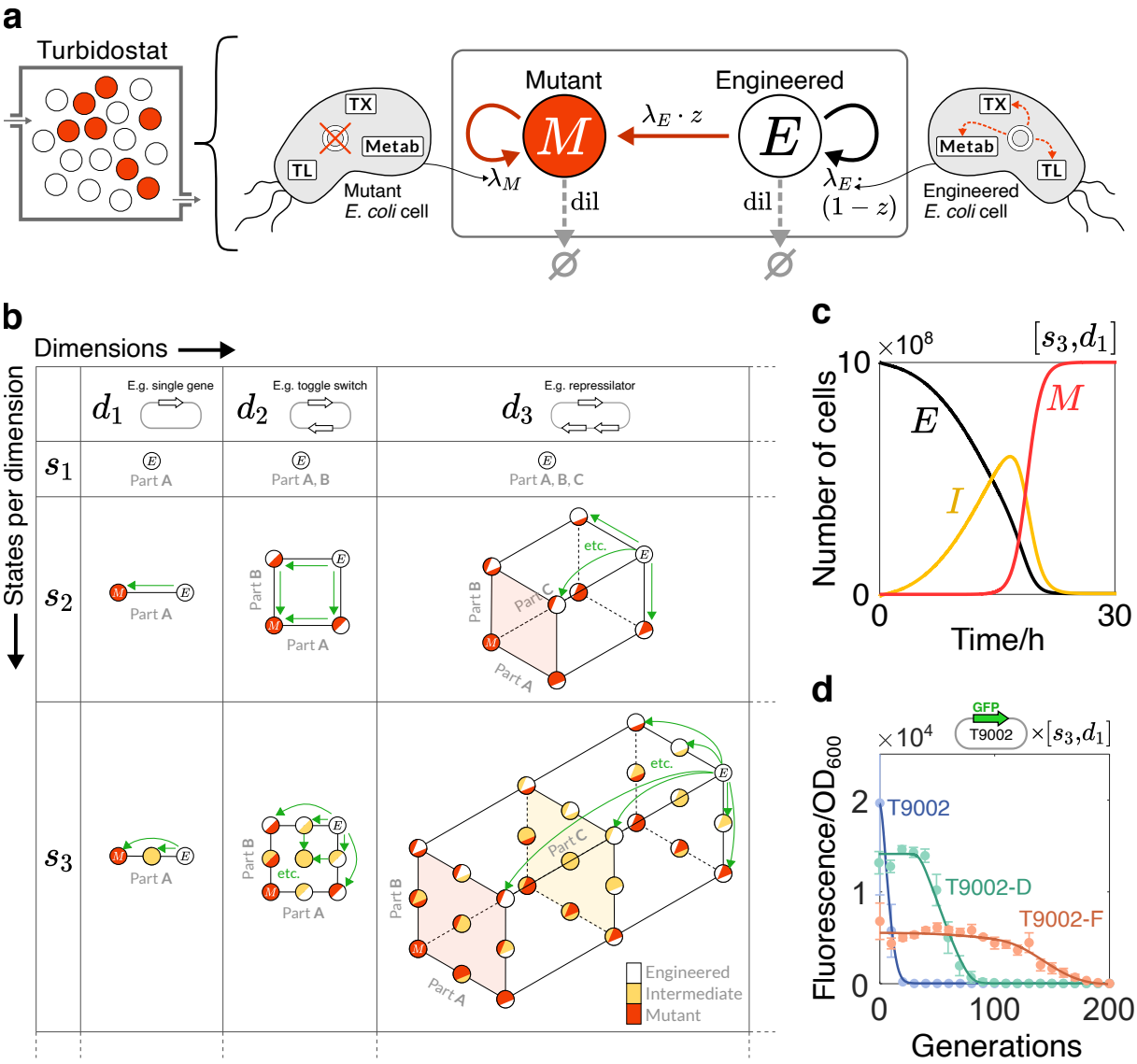

Figure 1: The modular mutation-aware framework. **(a)** The base model is represented by a two-state chemical reaction network in a turbidostat: engineered ('E') cells express synthetic genes while mutant ('M') cells do not. Cells belonging to a particular state are assumed to be identical. Transition rates are shown next to arrows. λ_E, λ_M: growth rate of E-cells and M-cells, respectively; z: probability that an E-cell produces one E-cell and one M-cell upon division; dil: dilution rate from the turbidostat. Growth rates are inferred using a host-aware model of *E. coli* that considers transcription ('TX'), translation ('TL') and energy metabolism ('Metab'). A different cell model is used for each state to reflect the difference in mutation phenotypes. **(b)** The two-state model is extended to consider variation in mutation location ('dimensions') and mutation severity ('states per dimension'). The dynamics of each mutable part are captured by states positioned along new orthogonal axes. Additional states added to a particular axis add variation in mutation severity, with states ordered in severity. For each state, a new cell model is used that reflects the state's mutation phenotype. Transitions between states (green arrows) occur independently and are mono-directional such that mutation severity can only increase. d: number of dimensions. s: number of states per dimension. $[s_y,d_x]$ a framework where $[s,d] = [y,x]$. Any integer values for s and d are permitted. **(c)** Simulation of a $[s_3,d_1]$ framework where a single gene's promoter is considered to vary between three mutation states: E, I (intermediate) and M. The number of cells belonging to each state is plotted over time. Initial conditions: $[E,I,M] = [10^9,0,0]$. Parameter values: $[\alpha_E, z_M, \alpha_I, z_I] = [10^4, 10^{-9}, 6{\times}10^3, 10^{-2}]$. **(d)** Simulation results from the $[s_3,d_1]$ framework are fit to fluorescence decay data from [14]. Each curve represents a variant of the synthetic construct 'T9002' grown in *E. coli* cells. For simulations, 'fluorescence' is the total amount of synthetic protein currently in the population and 'time' is converted to 'generations' using the cells' growth rates. See SI (Section S2) for parameter values and fit statistics.

We integrate Weisse et al.'s framework into our mutation model by forming one set of ODEs for each cell state. A given set of ODEs contains an equation for that state's cell quantity (such as Equation (2) for E-cells) alongside the host-aware cell equations (denoted as a subset of equations called 'Ψ', SI Section S1.1) from [22]. Within this structure, each equation set is adjusted to reflect its associated state's mutation phenotype by considering specific parameter values for the synthetic construct's expression dynamics.

One such parameter is the maximum transcription rate (α) which can be interpreted as the construct's promoter strength. By setting α to 0, for example, transcription is fully inhibited and the synthetic gene's expression becomes inactivated, hence capturing the behaviour of an M-cell. To model the dynamics of the whole population, each set of ODEs for all cell states are simulated together. By letting $\dot{\Psi}_X$ denote the host-aware ODEs specific to state X, the full set of ODEs composing our two-state mutation-aware model is defined as:

$$\text{Two-state mutation-aware model} \triangleq \{\dot{E}, \dot{\Psi}_E, \dot{M}, \dot{\Psi}_M\}. \tag{4}$$

2.2 Extending the framework to capture mutation heterogeneity

In reality, mutations to synthetic constructs can affect a cell's phenotype in different ways depending on (i) the functional part they affect (mutation 'location') and (ii) the degree to which they affect it by (mutation 'severity'). To capture these aspects of mutation heterogeneity, the two-state framework can be augmented by considering additional states that represent different mutation phenotypes (Figure 1b).

Location-specific mutations can be modelled by adding a dimension to the framework in Figure 1a for each new part being considered. We define a 'dimension' as a distinct functional unit, such as a promoter or an RBS. Fine-graining mutable elements into further sub-functions could be considered, such as splitting a promoter into '-35', 'spacer' and '-10' regions, however these details would need to be captured in the corresponding cell model equations. In order to analyse the mutation dynamics of x parts simultaneously, x orthogonal axes are drawn whereby the states along one axis correspond to one part's mutation state. As such, moving parallel to an axis represents a change to the mutation phenotype of that axis' associated part. We define the parameter d as the number of dimensions in a framework, with the notation d_x denoting that $d = x$, such as column 'd_2' in Figure 1b representing frameworks that model mutations to

two parts. As depicted in Figure 1b, dimensions can also represent mutations to parts across multi-gene constructs: d_3, for example, could capture the mutation dynamics of a three-gene repressilator by considering each gene's promoter for mutation. We assume that mutations to different parts develop independently, such that transitions (green arrows in Figure 1b) are restricted to one axis at a time. Furthermore, we assume that each part functions independently, such that the occurrence of one mutation does not influence the probability of another. Finally, we assume that mutations cannot be reversed, creating a one-way funnel of transitions from the E-state to M-state.

Variation in mutation severity can be included via additional states along each axis. Cells within these intermediate ('I') states are modelled using parameter values that reduce the activity of the corresponding part, but do not completely inactivate it. For a gene's promoter, for example, the maximum transcription rate of cells within an I-state would satisfy $\alpha_E > \alpha_I > 0$, where α_X denotes the value of α within state X. Less impactful mutations are assumed to occur with a distinct probability, requiring different parameters for the probabilities of creating I-cells (z_I) and M-cells (z_M) upon division. Additional I-states can be added to each axis to represent states with varying degrees of mutation severity, with these states being ordered by decreasing part functionality. Furthermore, we allow parts to sustain any mutation that reduces the part's functionality, meaning that individual transitions can bypass I-states. Akin to the parameter d, we define the parameter s as the number of states per dimension, with the notation s_y denoting that $s = y$. For convenience, we use the notation $[s_y,d_x]$ to represent a framework with x dimensions and y states per dimension. Frameworks up to $[s_3,d_3]$ are shown in Figure 1b, however our model generates the appropriate framework structure for any integer values of s and d (method outlined in SI Section S1.2).

A complete set of equations describing any framework can be represented by generalising Equation (4). For a mutation state X_i where $i \in [1, n]$, the full set of ODEs composing our mutation-aware model is defined as:

$$\text{General mutation-aware model} \triangleq \{\dot{X}_1, \dot{\Psi}_{X_1}, \ldots, \dot{X}_n, \dot{\Psi}_{X_n}\}. \tag{5}$$

As before, the synthetic construct's expression parameters in a particular state, such as the maximum transcription rate α, are chosen to reflect the mutation phenotype of the cells in that state.

2.3 Visualising state transitions and fitting to experimental data

To illustrate typical changes between states over time, a simulation of the $[s_3,d_1]$ framework is shown in Figure 1c where a single gene's promoter is considered to vary between three mutation states. Here, all cells start in the E-state (black line), before transitioning to I- and M-states (yellow and red lines, respectively). In a two-state population model, the E-state and M-state curves would be expected to have symmetric rates of change, as any cell leaving the E-state becomes part of the M-state by definition. The presence of an I-state, however, skews these curves, most notably seen by the E-state curve having an inflection point below its half-maximal value. This added asymmetry from 'partial' mutations may be important when capturing a range of mutation dynamics in populations.

To demonstrate the flexibility of our model in capturing mutation dynamics, and in turn to show the use of capturing mutation heterogeneity, we use our $[s_3, d_1]$ framework to fit simulation results to experimental data from Sleight et al. [14] (Figure 1d). As in Figure 1c, we apply this framework by modelling a single gene's promoter varying between three mutation states in order to approximate the effect of mutations that fully inactivate the construct. In their experiments, the authors engineer *E. coli* with variants of the same fluorescent-tagged synthetic construct ('T9002' and six variants appended with '-A' to '-F'), and record how each population's fluorescence decays over time due to the onset of mutations. Fits to three of these experiments are shown here (T9002 in blue, T9002-D in green and T9002-F in orange), with more complete details of fitting to all experiments given in the SI (Section S2). In our simulations, 'fluorescence' represents the total amount of synthetic protein currently in the population, and 'time' is converted to 'generations' using the cells' growth rates. The full details of these adjustments are given in the SI (Section S2).

The different designs associated with each T9002 construct cause changes to both the mutation probability of the construct and the associated growth rate of the cells. The precise combination of these factors is typically unknown without extensive experimental investigation, however they can be estimated by fitting key mutation parameters in our model $(\alpha_E, z_M, \alpha_I, z_I)$ to the experimental data. For the experiments shown in Figure 1d, optimal fits are obtained for T9002 and T9002-F using values for α_I and z_I, while fitting to T9002-D was best achieved without these parameters. This shows how our model can be used to describe mutation dynamics with flexibility, while suggesting a role for mutation heterogeneity in some scenarios. A more

complete dissection of these fits, alongside a discussion of the validity of the parameters values, is given in SI Section S2.

3 Predicting the yield and genetic stability of synthetic gene constructs

The impact of mutations on protein production is of particular interest in biotechnology, affecting both product yields and reliability of experiments [1, 26]. While cell strains with higher expression loads produce higher synthetic protein yields per unit time, they are also selected against faster in a growing cell population [14]. This trade-off between genetic shelf life and protein production yield is often unclear, and can be explored using our mutation-aware framework.

3.1 Quantitative metrics for synthetic protein production

In order to evaluate protein production through the lens of mutations, quantitative metrics can be defined in the context of our model. In a turbidostat, the yield of a given protein type can be defined as the cumulative amount of that protein extracted from cells that leave the chamber as a result of dilution. Here-in, 'heterologous' (H) will be used to describe proteins and variables related to the expression of synthetic gene constructs. The rate of H-production attributed to one cell type 'X_i' can be calculated by multiplying the amount of H-protein in one cell ($H_{X_i}(t)$), its growth rate ($\lambda_{X_i}(t)$), and the abundance of that cell type in the turbidostat ($X_i(t)$). The rate of H-production in the entire population is therefore the sum over all cell types. To compare results between continuous cultures of different sizes, we can furthermore represent this quantity as a per-cell average by dividing by N. It follows that the per-cell average rate of H-protein production in the population (unit: $\text{molecs}\,\text{h}^{-1}\,\text{cell}^{-1}$) is:

$$H_{\text{rate}}(t) = \sum_{i=1}^{n} \frac{H_{X_i}(t) \cdot \lambda_{X_i}(t) \cdot X_i(t)}{N \cdot \ln(2)}, \tag{6}$$

with the factor $\ln(2)$ required to link growth rate to doubling time. In turn, the per-cell average protein yield (unit: $\text{molecs}\,\text{cell}^{-1}$) between two time points T_0 and T is the integral of $H_{\text{rate}}(t)$:

$$H_{\text{yield}}(T_0, T) = \int_{T_0}^{T} H_{\text{rate}}(t)\ dt. \tag{7}$$

Intuitively, Equation (7) represents the amount of new cell mass that is converted to H-protein as the population grows during the time interval $[T_0, T]$.

Figure 2a shows how $H_{\text{yield}}(T_0, T)$ can be obtained from our framework in a simple case. Here we consider a population of cells expressing a single synthetic construct whose promoter is subject to fully-inactivating mutations ($[s_2,d_1]$). As t approaches 30 h in this simulation, the accumulation of non-producing M-cells leads to $H_{\text{rate}}(t)$ decreasing to zero, meaning that no additional protein yield can be obtained.

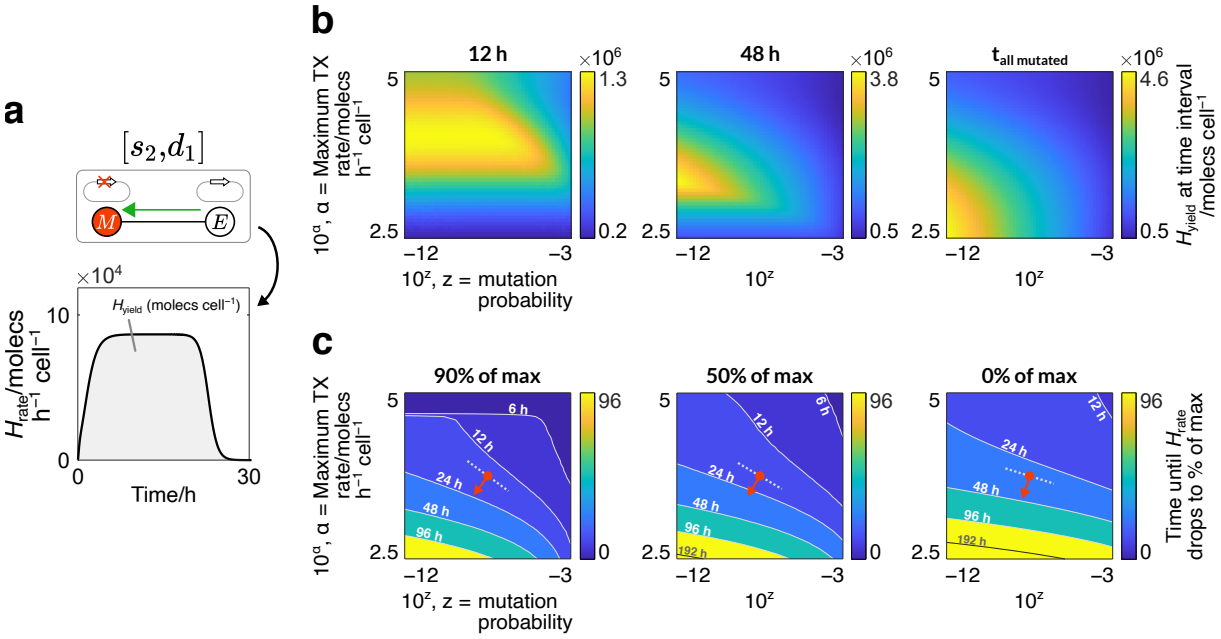

Figure 2: Predicting protein yield and viability of protein production. **(a)** Simulation of a framework where fully-inactivating mutations of a single gene's promoter ($[s_2,d_1]$) are modelled. The rate of synthetic protein production ($H_{\text{rate}}(t)$) is plotted. The synthetic protein yield ($H_{\text{yield}}(0, T)$) is the area under this curve. Parameter values: $[\alpha, z] = [1.05\times10^4\,\text{molecs}\,\text{h}^{-1}\,\text{cell}^{-1}, 10^{-12}]$. **(b)** Heat maps of synthetic protein yield for the $[s_2,d_1]$ framework in (a), calculated by varying the mutation probability (z) and the maximum transcription rate (α). Yield measurements are taken at three three intervals ($[0, T]$ where $T \in \{12\,\text{h}, 48\,\text{h}, t_{\text{all mutated}}\}$) to show how the optimal parameter combinations change over time. **(c)** Contour maps of the time taken for $H_{\text{rate}}(t)$ to drop to certain percentages (90%, 50% and 0%) of its maximum ('max') value for the $[s_2,d_1]$ framework in (a), calculated by varying z and α as in (b). Tracing any contour line gives the combinations of z and α that ensure a minimum protein production rate (given by the map) for the duration of that contour's time value.

3.2 Evaluating the effect of synthetic construct design on synthetic protein expression

When growing a cell population for synthetic protein extraction, it may not be desirable to wait until the mutations have become widespread before altering or terminating the experiment. This is because (i) beyond a certain time, the quantity of experimental resources required may outweigh the gains in H_{yield}, and (ii) researchers may want to maintain a population above a particular

threshold of active synthetic gene expression by restocking the turbidostat chamber with viable cells. For these reasons, it is useful to simulate the impact of different construct designs on protein production when measuring (i) the yield at fixed time intervals, and (ii) the time taken for the population's protein expression rate to drop below a particular threshold. These can be calculated by varying parameters that are intimately tied to the construct's design: the mutation probability (z), which is significantly influenced by a construct's sequence composition [10], and the maximum transcription rate (α), which depends on the choice of promoter and affects the cell's growth rate via the diversion of shared cellular resources towards heterologous gene expression rather than biomass production.

3.2.1 Predicting synthetic protein yield

While it is intuitive that increasing z decreases H_{yield}, the effect of increasing α is less obvious as higher expression rates typically correlate with slower growth rates. To explore this further, we calculate $H_{\text{yield}}(0, T)$ for the simple $[s_2,d_1]$ framework in Figure 2a over different intervals ($T \in \{12\,\text{h}, 48\,\text{h}, t_{\text{all mutated}}\}$), and observe how the optimum combinations of z and α are predicted to vary (Figure 2b). Here, $t_{\text{all mutated}}$ is the time at which all cells in the turbidostat have become M-cells. Results are displayed over time as successive heat maps, with lighter regions indicating combinations of z and α that produce more H-protein. As time progresses the optimum value for α drops, suggesting that lower protein expression delivers higher H_{yield} when culturing cells over longer time periods. This can be explained via the trade-off between the rate of protein expression and the effect of selection pressure: while higher transcription rates deliver more H-protein per unit time, these cells use more shared cellular resources and so grow more slowly compared to those with weaker H-protein expression. They are therefore diluted from the turbidostat at a faster rate, leaving fewer protein-producing cells in the population. This adverse effect on protein yield is relatively small at the beginning of an experiment but becomes more significant as time progresses, leading to a steady decline in the optimal value of α.

For the mutation probability z, the range of values that maximise protein yield becomes smaller over time, with later time points more significantly penalising higher values of z. This can be explained by considering that the accumulation of mutant cell types in a population accelerates over time. During early population growth, a limited number of mutant cells are produced according to the mutation probability, however as these cells grow faster than other cell types, a tipping point is eventually reached that allows their combined growth to accelerate beyond

the growth of other cell types. This aligns with the observed protein production dynamics from Figure 1d and Figure 2a where a fast decline in protein production is typically seen following a sustained period of protein expression. Together, these results suggest that optimising a synthetic construct's sequence for mutation becomes more crucial the longer that protein production is sustained for.

3.2.2 Predicting genetic shelf life

A separate consideration is monitoring a population's rate of protein production over time. This may allow researchers to gain insight into their device's 'genetic shelf life' and in turn allows them to predict when a system should be replenished with unmutated (E-state) cells. The impact of synthetic construct design on these aspects can be analysed using contour maps of the time taken for the population's protein expression rate ($H_{\text{rate}}(t)$) to drop to certain percentages of its maximum value, $\max(H_{\text{rate}}(t))$ (Figure 2c). In each map, tracing any contour line gives the combinations of z and α that ensure a minimum protein production rate (given by the map) for the duration of that contour's time value. For example, if a user wants to design their construct to sustain a minimum of 50% production rate for at least 24 h, they could choose any parameter values bounded below the '24 h' contour line in the middle map.

In each contour map, it can be seen that higher z and α lead to faster mutation accumulation, as displayed by the upper-right regions of the maps displaying contour lines with lower time values. This is as expected, as both of these parameters are positively associated with mutation pressure. Other trends can be seen by comparing the efficacy of z and α on increasing shelf-life. To illustrate this, a red mark is displayed on each contour map that represents a synthetic construct with the same maximum transcription rate and mutation probability. The arrows from these marks point in the direction of 'maximum increase' in contour time values, such that changing α and z in this direction is the most direct way to increase the genetic shelf life. As more mutations are allowed to accumulate in a population (moving between maps left to right), the direction of maximum increase becomes more vertical, meaning that α has an increasingly large effect on shelf life relative to z. This can also be seen by the contour line associated with this mark (dashed white line) 'straightening out' between maps (left to right). These results suggests that, if researchers are committed to growing cells until cell viability drops to low values, they can better sustain synthetic gene expression by varying the cell's rate of protein production (e.g. by changing the promoter) rather than its mutation probability (e.g. by removing mutagenic

sequence features). While this trend is specific to the construct design indicated by the red mark, it can be seen to hold for many other designs as the contour lines straighten between maps (left to right) across large regions of the parameter space.

4 Mutation-aware insights into common synthetic gene regulatory networks

So far, we have shown how our framework cam be used to explore the mutation dynamics of a single-gene construct with a simple protein output. Although this is useful to capture fundamental trends in synthetic protein expression, it does not represent more complex genetic devices such as those with multiple genes and regulatory components. To better demonstrate our framework's potential, we apply it to two of the most influential gene regulatory motifs in synthetic biology: the toggle switch and the repressilator. The mutation dynamics and subsequent protein expression behaviours of these devices are unintuitive *a priori* and so constitute prime examples for the application of our mutation-aware modelling framework.

4.1 Genetic toggle switches lose their capacity for switching over time

In its simplest form, a toggle switch comprises two genes whose protein outputs mutually inhibit the transcription of the other. They are typically designed to be bistable, with the two asymptotically-stable steady states corresponding to one gene being highly expressed while the other is silenced and *vice versa*. Specific external protein inhibitors can be added at a sufficiently high concentration to 'switch' the toggle from one steady state to another. This capacity for switching is arguably the most fundamental feature of a toggle switch and is hence used in multiple applications, from the regulation of gene expression [27, 28] to controlling biofilm formation [29]. Until now, theoretical analyses of the switching behaviour of toggle switches have assumed unchanging protein expression dynamics [30–32], whereas in reality, one would expect mutation accumulation to reduce the ability of these genetic constructs to switch between states.

4.1.1 A metric for assessing switching capacity

In our analysis, we consider a symmetric toggle switch functioning in an *E. coli* chassis with mutually repressing proteins 'A' and 'B' at concentrations A and B respectively (Figure 3a).

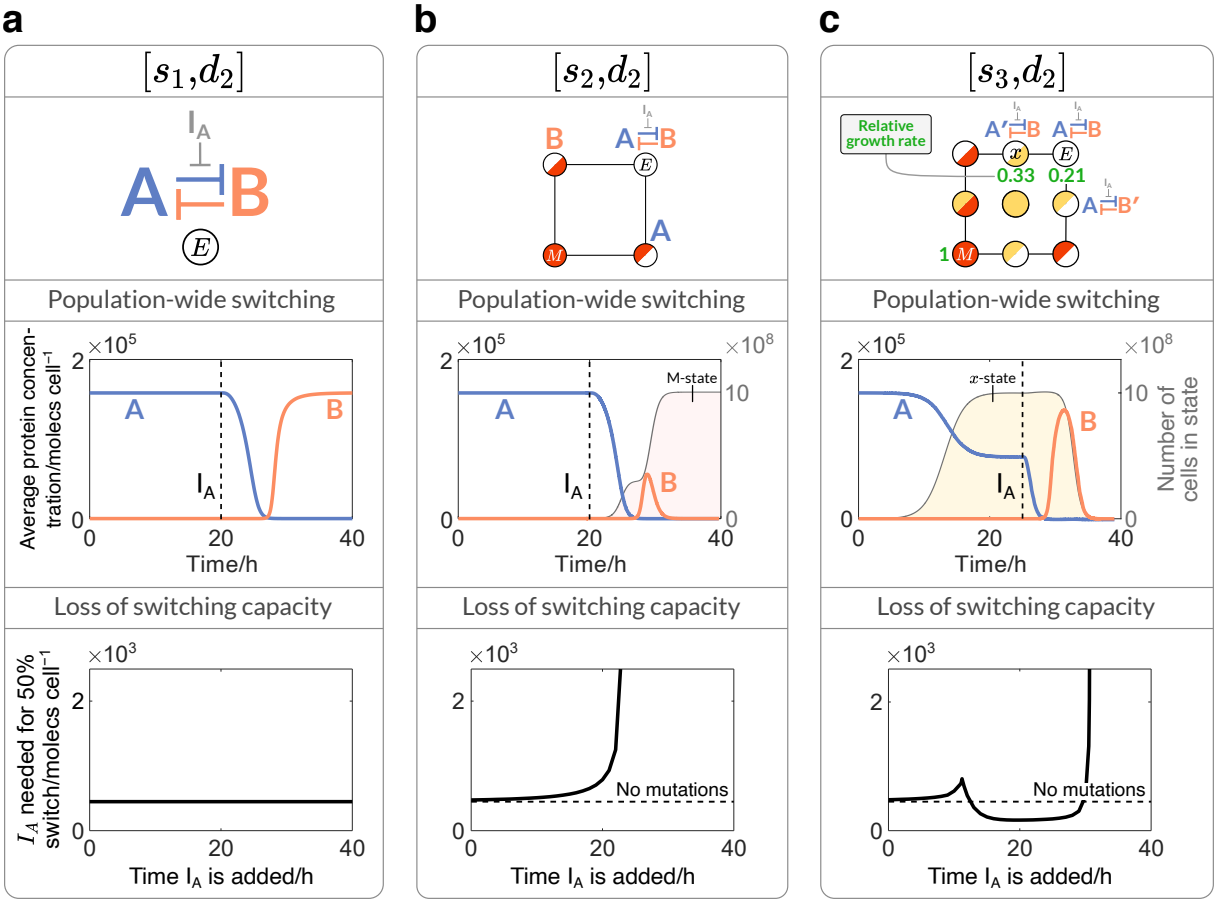

Figure 3: Applying our framework to study the mutation-driven dynamics of a genetic toggle switch. **(a)** Simulation of a cell population expressing a toggle switch without mutations ($[s_2,d_1]$). *Top*: the gene network topology of a toggle switch whose genes express proteins 'A' and 'B', and an inhibitor to the repression of B labelled 'I_A'. *Middle*: per-cell average concentrations of A and B over time. I_A is added at $t = 20\,\mathrm{h}$ (vertical dashed line), causing a transition from the 'high-A' steady state to the 'high-B' steady state. *Bottom*: the concentration of I_A required for at least 50% of the cells to transition from 'more-A' to 'more-B'. Variation in this quantity is monitored when adding inhibitor at different time points. For each gene, $[\alpha, z] = [10^4\,\mathrm{molecs\,h^{-1}\,cell^{-1}}, 0]$. **(b)** A simulation of a population expressing a toggle switch where fully-inactivating promoter mutations to both genes ($[s_2,d_2]$) are modelled. *Top*: each mutation state can be mapped onto the framework shown. All cells start in the non-mutated E-state (top-right). *Middle*: as in (a) for the left y-axis, with the right y-axis denoting the number of cells in the M-state (red area). I_A is added at $t = 20\,\mathrm{h}$ and is assumed to be mixed evenly in the population. *Bottom*: as in (a), with a dashed line showing the concentration of I_A required when no mutations occur. For each gene, $[\alpha, z] = [10^4\,\mathrm{molecs\,h^{-1}\,cell^{-1}}, 10^{-9}]$. **(c)** An equivalent system to (b) but additionally considering 'partial' mutations to each gene's promoter ($[s_3,d_2]$). *Top*: as in (b), with partially-mutated genes annotated with a prime ($'$) symbol. The growth rate of cells within select states relative to cells in the M-state is given as green numbers. *Middle*: as in (b), but with the right y-axis showing the number of cells in the upper-middle state (yellow area), and I_A being added at $t = 22\,\mathrm{h}$. *Bottom*: as in (b). For each gene, $[\alpha_E, z_M, \alpha_I, z_I] = [10^4\,\mathrm{molecs\,h^{-1}\,cell^{-1}}, 10^{-9}, 10^3\,\mathrm{molecs\,h^{-1}\,cell^{-1}}, 10^{-3}]$.

Our simulations start in the asymptotically stable steady state corresponding to 'high-A'. At a given time, a species I_A is added at concentration I_A to inhibit the ability of A to repress the expression of B. I_A is assumed to be evenly mixed in the population such that it is distributed to each cell type in proportion to its abundance. If I_A is sufficiently high, a switch from the stable high-A state to the stable high-B state occurs in the cells that still express synthetic proteins. As

time passes, however, the accumulation of mutations means that an equivalent concentration of inhibitor may lead to fewer cells 'switching'. In other words, the onset of mutations may cause the population to lose its capacity to switch from one stable state to the other over time. These effects are qualitatively illustrated in the middle row of Figure 3, which tracks the population-wide concentrations of A and B averaged per cell. If P_{X_i} denotes the concentration of a protein within a general cell type, and if the quantity of that cell type in the population is X_i, then we define the per-cell average concentration of that protein (P, unit: molecs cell^{-1}) as:

$$P = \sum_{i=1}^{n} \frac{P_{X_i} \cdot X_i}{N}. \tag{8}$$

To quantitatively track a population's change in switching capacity, we add $\mathrm{I_A}$ at successive time points and different concentrations, and record the number of cells whose protein concentrations satisfy $B > A$. To compare the effect of $\mathrm{I_A}$ across different simulations, at each time point of adding $\mathrm{I_A}$ we record its concentration required to induce at least 50% of the population's cells to switch from one steady state to another. It is noted that this metric does not consider the distance between fixed points in phase space, but only records the ability to switch between stable states. Our measure of a population's switching capacity is therefore defined purely as the ability to switch states rather than the 'strength' of the corresponding switch.

4.1.2 Impact of mutations on the switching capacity of genetic toggle switches

With no mutations, adding the inhibitor $\mathrm{I_A}$ at a sufficiently high concentration causes all cells in the population to 'switch', independent of when it is added. This can be seen from the constant value of I_A in Figure 3a and suggests that the switching capacity of the system does not change over time, a feature found in traditional theoretical analyses of toggle switches.

To explore the impact of mutations we can first apply the [s_2,d_2] framework, which models fully-inactivating promoter mutations to each synthetic gene (Figure 3b). The top panel shows how the construct design that corresponds to each mutation phenotype is mapped on to each state. When $\mathrm{I_A}$ is added to this system at the same concentration and at the same time point ($t = 20\,\mathrm{h}$) as in the 'no mutations' case, different dynamics are seen: the effect of inhibition starts a transition from 'A' to 'B', however due to the accumulation of mutations, the population-wide concentration of B crashes and so the number of cells with 'more-B' never surpasses 50%. The bottom panel shows that surpassing 50% is still possible when adding $\mathrm{I_A}$ at $t = 20\,\mathrm{h}$, but only

when using a higher concentration ($\approx 0.8\times10^3$ molecs cell^{-1}). The monotonic increase of this curve shows that the later I_A is added, the greater the concentration required to cause the same number of cells to 'switch', indicative of the greater onset of mutations. At some time point, mutations have accumulated enough that the number of switching cells never reaches 50%, as shown by the curve approaching a vertical asymptote. In essence, the increasing value of I_A is indicative of the system losing switching capacity over time, with it being completely lost at the time point of the graph's vertical asymptote.

As explored with our framework designs in Section 2.2, mutations may not completely inactivate a component, but could also partially reduce its activity. Such details may be important to capture when attempting to fully understand the dynamics of complex gene regulatory networks. These effects can be considered for the toggle switch by applying our [s_3,d_2] framework, which additionally models mutations that partially inactivate each gene's promoter (Figure 3c, top). The construct designs associated with select states are shown, with genes containing partially-mutated components being annotated with a prime ($'$) symbol. These 'partial mutations' cause genes to produce proteins at a lower rate, leading to reduced resource consumption and faster cell growth. This is indicated by the values adjacent to each cell state in the top panel of Figure 3c, showing the growth rate of cells in that state relative to M-cells. A consequence of cells with higher growth rates is that they become progressively dominant in a population. This can be seen in the middle panel where the number of cells in the state with partially-reduced expression of A (top-middle cell state in the top panel) is plotted over time. The presence of this mutation phenotype enables the population-wide concentration of A to restabilise at a new lower value after initially declining, indicative of transitions from the E-state to intermediate states.

The effect of different mutation phenotypes accumulating can change a system's switching capacity. In the [s_2,d_2] framework it was found that more inhibitor was required to induce the same degree of switching, however when inputting the amount of I_A at a later time point ($t = 22$ h) in the [s_3,d_2] framework, *more* cells switch from 'more-A' to 'more-B' (higher peak for 'B' in middle panel). This is due to the population-wide concentration of A stabilising at a lower value which in turn requires less I_A to induce switching. As such, graphing the concentration of I_A required for a 50% switch over time (lower panel) reveals that, while the onset of mutations initially causes the population's switching capacity to drop, it then starts to regenerate at $t \approx 11$ h. This reversal continues until the value of I_A falls below the level recorded in the 'no mutations' case (dashed line), suggesting that the presence of mutation heterogeneity can actually *increase* a

population's switching capacity during particular time intervals. It is noted that this only holds for some parameter values of α_I, as when the accumulation of I-cell phenotypes is lower, the system returns to a monotonic loss in switching capacity over time (SI Section S3.3).

4.2 Repressilators are resistant to single points of failure

Repressilators are genetic oscillators typically consisting of three genes inhibiting each other in a one-way ring structure (Figure 4a) with widespread application in synthetic biology [33]. Naturally-occurring genetic oscillators are robust to shifts in period or amplitude, often due to in-built feedback mechanisms [34, 35], however when implemented synthetically, their ability to sustain regularly-repeating oscillations diminishes [36, 37]. One root cause for a this loss of regularity may be the oscillator's synthetic its genes losing functionality over time, however experimental data that explores the role of mutations is scarce. We therefore apply our framework to examine a synthetic repressilator's dynamics in a growing cell population, and in turn explore how mutations may impact their regularity over time.

4.2.1 Modelling repressilators with no mutations

In what follows, we consider a repressilator composed of three genes that express proteins 'A', 'B' and 'C' at concentrations of A, B and C, respectively (Figure 4a, top). As before, we plot the per-cell average concentrations With no mutations, the construct behaves as expected with the per-cell average concentrations for each protein (Equation (8)) oscillating with uniform period and amplitude (middle panel). To capture the impact of circuit expression on the host, we additionally record the energy capacity (left y-axis of lower panel), defined as the per-cell average concentration of 'energy molecules' (SI Equation S2a, [22]). These energy molecules are essential for the expression of any gene in the cell and thus constitute a key growth-limiting resource. The energy capacity is plotted alongside the cell's growth rate relative to an M-cell (right y-axis of lower panel).

4.2.2 Effect of fully-inactivating mutations on oscillations

As with the toggle switch analysis, we can first apply a framework that models severe mutations to each synthetic gene's promoter (Figure 4b). The top panel shows how the construct design that corresponds to each mutation phenotype is mapped on to each state of the $[s_2,d_3]$ framework.

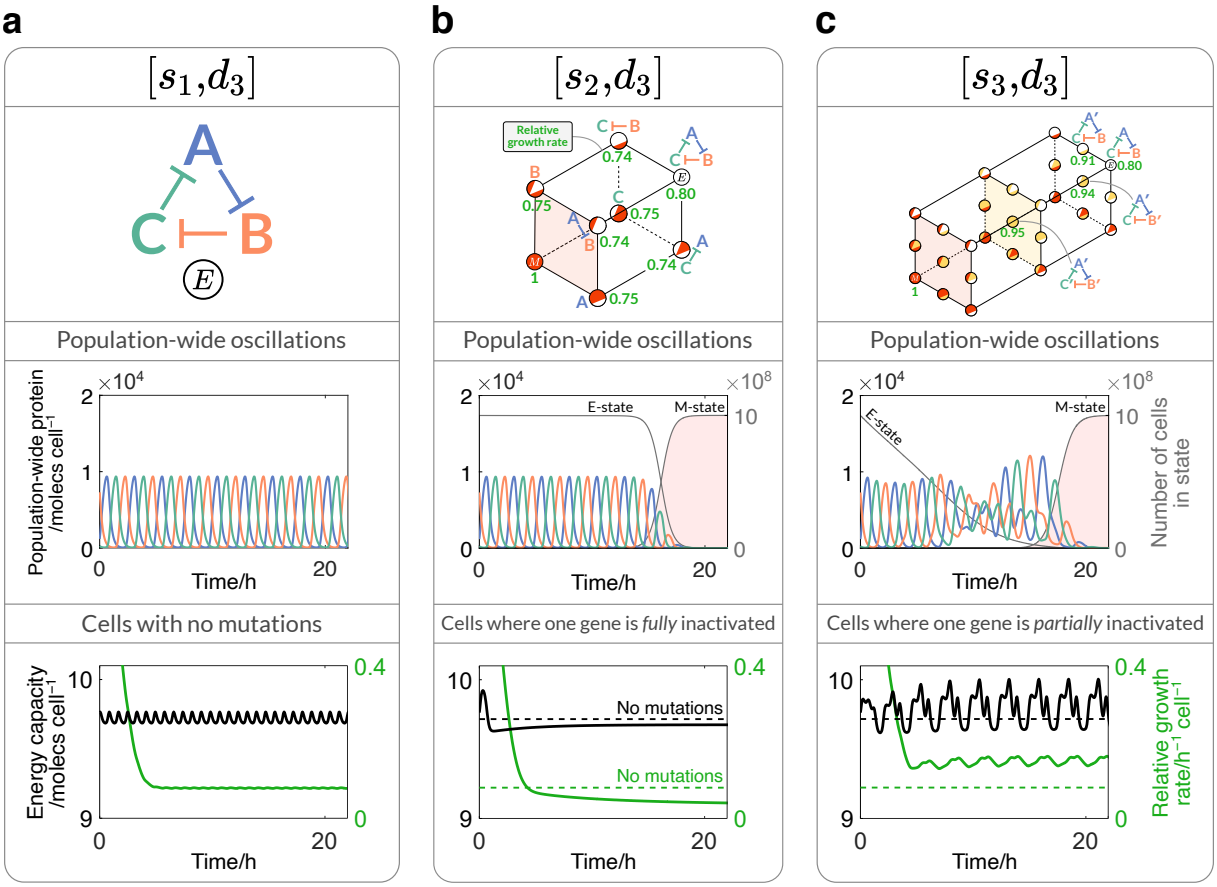

Figure 4: Applying our framework to study the mutation-driven dynamics of a repressilator. **(a)** Simulation of a cell population expressing a repressilator without mutations ($[s_1,d_3]$). *Top*: the gene network topology of a repressilator whose genes express proteins 'A', 'B' and 'C'. *Middle*: population-wide concentrations of A, B and C per cell over time. *Bottom*: energy capacity per cell (left y-axis, black) and growth rate per cell relative to M-cells (right y-axis, green). For each gene, $\alpha = 10^5\,\text{molecs}\,\text{h}^{-1}\,\text{cell}^{-1}$. **(b)** Simulation of a cell population expressing the genes of a repressilator when fully-inactivating promoter mutations ($[s_2,d_3]$) are considered. *Top*: each mutation state can be mapped onto the framework shown. All cells start in the non-mutated E-state (top-right). The growth rate of cells within each state relative to cells in the M-state is given as green numbers. *Middle*: as in (a) for the left y-axis, with the right y-axis denoting the number of cells in the E- and M-states (white and red areas respectively). *Bottom*: the per-cell energy capacity and growth rate in cells with one synthetic promoter fully inactivated. Horizontal dashed lines indicate the values of energy capacity and growth rate for cells with no mutations. For each synthetic gene: $[\alpha, z] = [10^5\,\text{molecs}\,\text{h}^{-1}\,\text{cell}^{-1}, 10^{-3}]$. **(c)** An equivalent system to (b), but additionally considering 'partial' mutations to each gene's promoter ($[s_3,d_3]$). *Top*: as in (b) with partially-mutated genes annotated with a prime ($'$) symbol. *Middle*: as in (b). *Bottom*: as in (b), except modelling cells where one gene has a partially-inactivated promoter. For each synthetic gene, $[\alpha_E, z_M, \alpha_I, z_I] = [10^5\,\text{molecs}\,\text{h}^{-1}\,\text{cell}^{-1}, 10^{-3}, 10^4\,\text{molecs}\,\text{h}^{-1}\,\text{cell}^{-1}, 5{\times}10^{-1}]$.

As before, values adjacent to cell states indicate the growth rate of cells within that state relative to M-cells. When recording the protein concentrations over time, we see that the oscillations remain robust for an extended period of time before their amplitude eventually reduces to zero (middle panel). This is also reflected by the number of cells in the E- and M-states (right y-axis, white and red areas), whose values only change significantly towards the end of the simulation time.

These sustained regular oscillations can be understood by comparing the energy capacity and growth rate of cells in different states (bottom panel). When one gene is fully inactivated, the energy capacity in the cell falls below that of cells with no mutations (black dashed line corresponding to the average of the oscillating value from Figure 4a). As a result, there is less energy capacity to sustain protein expression which causes the growth rate to drop compared to cells with no mutations (green dashed lined). Cells with one gene inactivated will therefore get removed from the turbidostat at a faster rate than E-cells, making it harder for them to accumulate in the population. In other words, while transitions from the E-state to directly connected states are possible, there is a low selective pressure to do so, explaining why stable oscillations persist over an extended period of time. As M-cells have the highest growth rate in the population, M-cells gradually become dominant relative to other cell types. As a result, a tipping point is eventually reached whereby M-cells out-compete other cell states.

The drop in growth rate upon inactivating one of the respressilator's genes is initially counter-intuitive as one may expect cells to gain more resources for growth-supporting processes. However a plausible explanation is that, for cells with no mutations, sustaining a cycle of mutually repressing species prevents any one protein from reaching high expression levels and thus prevents an excessive drain on cellular resources. Removing one component and thereby breaking the repressilator cycle allows one protein to become dominant and in turn cause a higher drain on resources than when all three genes are active and repressing each other.

4.2.3 Effect of partially-inactivating mutations on oscillations

Mutation dynamics can be analysed in greater detail by additionally considering the effects of 'partial' mutations to each gene's promoter. The top panel in Figure 4c shows how this can be achieved using the $[s_3,d_3]$ framework, with construct designs associated with select states being shown. In contrast to the effect of severe mutations genes, the introduction of partial mutations causes the regularity of the oscillations to change from the offset, supported by a steady decrease of E-cells from the start of the simulation. This behaviour can again be understood by comparing the energy capacity and growth rate between states. The bottom panel shows that, when mutation events partially inactivate the promoter of one gene, the cell's energy capacity oscillates at a higher average value than with no mutations. In turn, the greater availability of resources allows for expression of proteins at higher levels, which boosts the growth rate of partial mutant cells relative to E-cells. This suggests that reducing the repression of any protein,

while not totally inactivating it, is selectively advantageous and allows associated mutant cells to accumulate in the turbidostat at a faster rate. Cells in some of these intermediate mutant states exhibit a wide variety of oscillating protein expression patterns (SI, Section S4.2), explaining why their accumulation causes the population-wide oscillations to quickly lose uniformity.

To explain the higher energy capacity resulting from partially-mutated genes, the strength of repression within different gene constructs can again be considered. Partially reducing the expression of one gene reduces its drain on the cell's resources. However, its remaining activity prevents the other protein species from dominating to excessive levels. The resulting sum of repression is therefore sufficiently low such that the cell's available energy capacity increases. Contrasting this with the effects of fully-inactivating mutations, it can be suggested that synthetic repressilators are typically resistant to single points of failure.

5 Discussion

5.1 Experimental context is important when evaluating the capacity for protein production

For many systems in the biotechnology industry, the ability to predict protein yield and genetic shelf-life is vital in order to maximise profits and experimental reliability. Our analysis in Section 3 showed how the design of synthetic constructs impacts both yield and shelf-life via the maximum transcription rate α and the mutation probability z, both of which can be altered by changing a construct's promoter or its nucleotide sequence [10], respectively. When optimising yield, we showed how the time scale of experiments impacts the choice of α due to a trade-off between production capacity and selection pressure (Figure 2b), a trend that is supported by recent experimental studies [38]. When improving shelf-life, we see that for many parameter values, changing α becomes increasingly more impactful than changing z as more mutations accumulate (Figure 2c). Growing populations to high degrees of mutation is not typically viable, however these results may empower researchers to justify one experimental change to their synthetic construct over another. This becomes more important when either resources or time are limiting, as it could be unfeasible to both modify a construct's parts (to change its growth rate, for example via α) and to streamline its sequence (to change its mutation probability).

Our framework has clear utility with predicting protein production dynamics, however assigning accurate parameter values requires knowledge about the part components (for transcription rates) and nucleotide sequence (for mutation probabilities). This is relatively simple for frameworks with two mutation states per dimension, where the parts and sequence of the non-mutated construct can be analysed to obtain approximations for α_E and z_M. Frameworks with three or more mutation states per dimension, however, require information about the behaviour of partially-mutated components and the probabilities of their associated mutations (i.e. α_I and z_I for partially-mutated genetic parts such as promoters). In theory, these parameters could be inferred by running preliminary cell cultures for a prolonged time and sequencing the constructs to obtain the distribution of mutation types.

Another consideration for modelling accuracy is that expressing genes to high levels can cause problems that our model does not capture, such as the accumulation of toxic metabolites [39] or misfolded proteins [40, 41]. These aspects would likely lead to further drops in yield and shelf-life relative to the values predicted using our framework. One study [42] reports that toxic effects from expression occur when more than 30% of the proteome contains heterologous proteins which could act as a useful starting point, however different proteins likely relate to different thresholds of toxicity and so accurate modelling would require re-quantifying these values for each new construct.

5.2 A mutation-aware approach uncovers design paradigms for synthetic constructs with predictable genetic stability

To better demonstrate our framework's potential, we explored the utility of varying the number of dimensions (d) and the number of mutation states per dimension (s). In general, the choice of s and d was crucial in accurately capturing mutation dynamics of a system. For simple cases with low mutation complexity, it may be appropriate to only consider fully-inactivating mutations to a single part ($[s_2,d_1]$), however when more detail is required, the number of mutable parts or the variation in mutation severity may be increased.

As a first illustrative example, we use our framework to analyse the impact of mutations on the dynamics of genetic toggle switches. When permitting partially-inactivating mutations to a toggle switch's promoter regions, we found that its capacity to switch between stable states can temporarily increase (Figure 3). In these instances, less inhibitor is required to induce switching, which could have conflicting experimental implications. On one hand, switching

behaviour can be induced with less expenditure on laboratory resources, delivering a simple economic benefit. Despite this, a higher switching capacity may make toggle switches more prone to 'accidental' switching resulting from leaky gene expression. These ideas have not yet been experimentally tested, however their implications could be important for applications that require precise and predictable switching behaviour. As shown in Bothfeld et al. [43], for example, systems engineered to decouple cell growth from synthetic product formation, need to precisely and reliably switch gene expression to manage cell toxicity and not waste product formation.

In our mutation-based analysis of repressilators, we predict that they are resistant to single points of failure such that such that completely inactivating any gene is selectively disadvantageous (Figure 4). As with the toggle switch, this idea has not been experimentally tested, however some studies examine the robustness of repressilator dynamics through other means. Potvin et al., for example, experimentally show that decreasing sources of noise in their repressilator designs improved the regularity of oscillations [44], while Qiao et al. [45] show that oscillation-inducing network designs improve robustness when they contain positive autoregulation. As a result, while our analysis suggests one source of irregularity in oscillations (i.e. through mutation heterogeneity), we note that this might not the sole determining factor. Whether through reducing noise or restricting sources of mutation, sustaining regular oscillations over extended time periods of time would aid many oscillation-driven applications, such as building genetic clocks [46] or cardiac pacemakers [47].

The framework applications above show how new paradigms for the engineering of synthetic constructs could be used to enhance future designs of many synthetic devices. The modularity of our framework means that these analyses can be extended to any gene regulatory network, potentially uncovering other unique insights into their evolutionary dynamics. Coherent and incoherent feedforward loops, for example, use combinations of activation and repression loops that make their resulting mutation dynamics not obvious *a priori* [48], and would therefore constitute interesting candidates to analyse via our framework in the future. As more focus is being directed to understanding and shaping the evolutionary routes a synthetic construct construct can take [3, 6], mutation-aware frameworks such as those introduced here will likely become crucial to the Design-Build-Test-Learn cycle of synthetic biology in the coming years.

5.3 Modelling capabilities can be enhanced by augmenting resource-based cell models

During the design of our mutation-aware framework, we made a series of modelling decisions to enable analysis of a range of synthetic construct designs. Most notably, we used the host-aware model from Weisse et al. [22] to calculate a value for cell growth rate: this model captures how the distribution of shared cellular resources impacts protein production rate when synthetic genes are expressed in an *E. coli* chassis. This resource-aware model is simple in comparison to other models [23–25], making it an ideal candidate to integrate into a larger framework like ours. However, its lack of granularity introduces some limitations when linking the design of synthetic constructs with cell growth and selection pressures.

Weisse et al.'s framework allows its users to change the values of parameters that correlate with a gene's maximum transcription rate and mRNA-ribosome binding rate, two key components of synthetic construct design. Despite this, it does not describe other core components such as the coding sequence (CDS), terminator and origin of replication (*ori*), and therefore cannot capture the effects of codon efficiency, termination efficiency, and copy number, respectively. For example, modelling the CDS in more detail would allow further insight into translation, which Weisse et al. only consider as a one-step process. This could be important when considering variable codon design, as inefficient codons can cause ribosomes to slow down and form ribosomal traffic jams as they translate mRNA transcripts [49–51], thereby becoming unusable for other cell processes. Many biophysical models of translation consider this [51–54], with Algar et al. [51] and Sarvari et al. [54] specifically modelling the impact of slow codons on shared cellular resources and growth rate. Adding these components to Weisse et al.'s model could add predictive power to the tool, particularly in scenarios where ribosomes become a limiting resource.

Finally, while our model is parameterised based on *E. coli*, it could be adapted to other bacterial hosts by using alternative parameter values. This would be useful when applying resource constraints to organisms that have different industrial applications, for example *Pseudomonas* [55, 56] and *M. genetalium* [57]. Modelling eukaryotic chassis may be more difficult due to their added complexity, however a recent model by Frei et al. [58] showcases a predictive cell model that successfully captures resource allocation in mammalian cells. Furthermore, other studies [59, 60] suggest that eukaryotic cells obey similar 'growth laws' [61] to bacteria, such as the proportionality between a cell's ribosomal protein fraction and its growth rate. Together, these

results suggest that extending our mutation-aware framework to a wider range of cell models is feasible.

6 Conclusion

In this paper, we have introduced a mathematical framework that enables users to easily model the mutation-driven dynamics inherent to engineered cell populations. We demonstrated a number of core use-cases, from predicting protein yields and genetic shelf-lives in growing populations, to uncovering new design paradigms of multi-gene circuits. Beyond individual applications, our results demonstrate that our framework is particularly suited to present-day synthetic biology, where increasingly complex experiments and varied construct designs are commonplace. As more focus is being given to understanding and shaping the 'evolutionary potential' of synthetic constructs, analysing these through the lens of mutations will become increasingly important, cementing our framework's utility in the coming years.

References

[1] Peter Rugbjerg and Morten OA Sommer. Overcoming genetic heterogeneity in industrial fermentations. *Nature biotechnology*, 37(8):869–876, 2019.

[2] Marko Ahteensuu. Synthetic biology, genome editing, and the risk of bioterrorism. *Science and engineering ethics*, 23(6):1541–1561, 2017.

[3] Simeon D Castle, Claire S Grierson, and Thomas E Gorochowski. Towards an engineering theory of evolution. *Nature Communications*, 12(1):1–12, 2021.

[4] Marta Ciechonska, Marc Sturrock, Alice Grob, Gerald Larrouy-Maumus, Vahid Shahrezaei, and Mark Isalan. Ohm's law for emergent gene expression under fitness pressure. *bioRxiv*, page 693234, 2019.

[5] Alan Costello and Ahmed H Badran. Synthetic biological circuits within an orthogonal central dogma. *Trends in biotechnology*, 39(1):59–71, 2021.

[6] Fernanda Pinheiro, Omar Warsi, Dan I Andersson, and Michael Lässig. Metabolic fitness landscapes predict the evolution of antibiotic resistance. *Nature Ecology & Evolution*, 5(5): 677–687, 2021.

[7] James J Truglio, Deborah L Croteau, Bennett Van Houten, and Caroline Kisker. Prokaryotic nucleotide excision repair: the uvrabc system. *Chemical reviews*, 106(2):233–252, 2006.

[8] Malgorzata Bzymek and Susan T Lovett. Instability of repetitive dna sequences: the role of replication in multiple mechanisms. *Proceedings of the National Academy of Sciences*, 98 (15):8319–8325, 2001.

[9] Karl Näslund, Peter Saetre, Jenny von Salomé, Tomas F Bergström, Niclas Jareborg, and Elena Jazin. Genome-wide prediction of human vntrs. *Genomics*, 85(1):24–35, 2005.

[10] Benjamin R Jack, Sean P Leonard, Dennis M Mishler, Brian A Renda, Dacia Leon, Gabriel A Suárez, and Jeffrey E Barrick. Predicting the genetic stability of engineered dna sequences with the efm calculator. *ACS synthetic biology*, 4(8):939–943, 2015.

[11] Itamar Menuhin-Gruman, Matan Arbel, Niv Amitay, Karin Sionov, Doron Naki, Itai Katzir, Omer Edgar, Shaked Bergman, and Tamir Tuller. Evolutionary stability optimizer (eso): A novel approach to identify and avoid mutational hotspots in dna sequences while maintaining high expression levels. *ACS Synthetic Biology*, 2021.

[12] Peter Rugbjerg, Nils Myling-Petersen, Andreas Porse, Kira Sarup-Lytzen, and Morten OA Sommer. Diverse genetic error modes constrain large-scale bio-based production. *Nature communications*, 9(1):1–14, 2018.

[13] Scott L Nuismer, Nathan C. Layman, Alec J Redwood, Baca Chan, and James J Bull. Methods for measuring the evolutionary stability of engineered genomes to improve their longevity. *Synthetic Biology*, 6(1):ysab018, 2021.

[14] Sean C Sleight, Bryan A Bartley, Jane A Lieviant, and Herbert M Sauro. Designing and engineering evolutionary robust genetic circuits. *Journal of biological engineering*, 4(1):1–20, 2010.

[15] Ruben Perez-Carrasco, Chris P Barnes, Yolanda Schaerli, Mark Isalan, James Briscoe, and Karen M Page. Combining a toggle switch and a repressilator within the ac-dc circuit generates distinct dynamical behaviors. *Cell systems*, 6(4):521–530, 2018.

[16] Garrett W Roell, Jian Zha, Rhiannon R Carr, Mattheos A Koffas, Stephen S Fong, and Yinjie J Tang. Engineering microbial consortia by division of labor. *Microbial cell factories*, 18(1):1–11, 2019.

[17] Rebecca A Burrell, Nicholas McGranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345, 2013.

[18] Marc J Williams, Benjamin Werner, Timon Heide, Christina Curtis, Chris P Barnes, Andrea Sottoriva, and Trevor A Graham. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature genetics*, 50(6):895–903, 2018.

[19] Stefano Cardinale, Marcin Pawel Joachimiak, and Adam Paul Arkin. Effects of genetic variation on the e. coli host-circuit interface. *Cell reports*, 4(2):231–237, 2013.

[20] Olivier Borkowski, Francesca Ceroni, Guy-Bart Stan, and Tom Ellis. Overloaded and stressed: whole-cell considerations for bacterial synthetic biology. *Current opinion in microbiology*, 33:123–130, 2016.

[21] Alice Boo, Tom Ellis, and Guy-Bart Stan. Host-aware synthetic biology. *Current Opinion in Systems Biology*, 14:66–72, 2019.

[22] Andrea Y Weiße, Diego A Oyarzún, Vincent Danos, and Peter S Swain. Mechanistic links between cellular trade-offs, gene expression, and growth. *Proceedings of the National Academy of Sciences*, 112(9):E1038–E1047, 2015.

[23] Nils Giordano, Francis Mairet, Jean-Luc Gouzé, Johannes Geiselmann, and Hidde de Jong. Dynamical allocation of cellular resources as an optimal control problem: novel insights into microbial growth strategies. *PLoS computational biology*, 12(3):e1004802, 2016.

[24] Chen Liao, Andrew E Blanchard, and Ting Lu. An integrative circuit–host modelling framework for predicting synthetic gene network behaviours. *Nature microbiology*, 2(12): 1658, 2017.

[25] David W Erickson, Severin J Schink, Vadim Patsalo, James R Williamson, Ulrich Gerland, and Terence Hwa. A global resource allocation strategy governs growth transition kinetics of escherichia coli. *Nature*, 551(7678):119, 2017.

[26] Francesca Ceroni, Alice Boo, Simone Furini, Thomas E Gorochowski, Olivier Borkowski, Yaseen N Ladak, Ali R Awan, Charlie Gilbert, Guy-Bart Stan, and Tom Ellis. Burden-driven feedback control of gene expression. *Nature methods*, 15(5):387, 2018.

[27] Beat P Kramer, Alessandro Usseglio Viretta, Marie Daoud-El Baba, Dominique Aubel, Wilfried Weber, and Martin Fussenegger. An engineered epigenetic transgene switch in mammalian cells. *Nature biotechnology*, 22(7):867–870, 2004.

[28] Tara L Deans, Charles R Cantor, and James J Collins. A tunable genetic switch based on rnai and repressor proteins for regulating gene expression in mammalian cells. *Cell*, 130(2): 363–372, 2007.

[29] Hideki Kobayashi, Mads Kaern, Michihiro Araki, Kristy Chung, Timothy S Gardner, Charles R Cantor, and James J Collins. Programmable cells: interfacing natural and engineered gene networks. *Proceedings of the National Academy of Sciences*, 101(22): 8414–8419, 2004.

[30] Timothy S Gardner, Charles R Cantor, and James J Collins. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767):339–342, 2000.

[31] Patrick Hillenbrand, Georg Fritz, and Ulrich Gerland. Biological signal processing with a genetic toggle switch. *PloS one*, 8(7):e68345, 2013.

[32] Kaushik Raj, William TZ Wong, Beini Zhang, and Radhakrishnan Mahadevan. Model guided design of enhanced bi-stable controllers to effectively switch cellular states. *bioRxiv*, 2022.

[33] Oliver Purcell, Nigel J Savery, Claire S Grierson, and Mario Di Bernardo. A comparative analysis of synthetic genetic oscillators. *Journal of The Royal Society Interface*, 7(52): 1503–1524, 2010.

[34] Albert Goldbeter et al. Biochemical oscillations and cellular rhythms. *Biochemical Oscillations and Cellular Rhythms*, 1997.

[35] Shu-Wen Teng, Shankar Mukherji, Jeffrey R Moffitt, Sophie De Buyl, and Erin K O'shea. Robust circadian oscillations in growing cyanobacteria require transcriptional feedback. *Science*, 340(6133):737–740, 2013.

[36] Michael B Elowitz and Stanislas Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.

[37] Mariette R Atkinson, Michael A Savageau, Jesse T Myers, and Alexander J Ninfa. Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in escherichia coli. *Cell*, 113(5):597–607, 2003.

[38] Francesca Ceroni, Rhys Algar, Guy-Bart Stan, and Tom Ellis. Quantifying cellular capacity identifies gene expression designs with reduced burden. *Nature methods*, 12(5):415, 2015.

[39] Sónia Carneiro, Eugénio C Ferreira, and Isabel Rocha. Metabolic responses to recombinant bioprocesses in escherichia coli. *Journal of biotechnology*, 164(3):396–408, 2013.

[40] Antonio Villaverde and M Mar Carrió. Protein aggregation in recombinant bacteria: biological role of inclusion bodies. *Biotechnology letters*, 25(17):1385–1395, 2003.

[41] Kerry A Geiler-Samerotte, Michael F Dion, Bogdan A Budnik, Stephanie M Wang, Daniel L Hartl, and D Allan Drummond. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proceedings of the National Academy of Sciences*, 108(2):680–685, 2011.

[42] Moshe Kafri, Eyal Metzl-Raz, Ghil Jona, and Naama Barkai. The cost of protein production. *Cell reports*, 14(1):22–31, 2016.

[43] William Bothfeld, Grace Kapov, and Keith EJ Tyo. A glucose-sensing toggle switch for autonomous, high productivity genetic control. *ACS synthetic biology*, 6(7):1296–1304, 2017.

[44] Laurent Potvin-Trottier, Nathan D Lord, Glenn Vinnicombe, and Johan Paulsson. Synchronous long-term oscillations in a synthetic gene circuit. *Nature*, 538(7626):514–517, 2016.

[45] Lingxia Qiao, Zhi-Bo Zhang, Wei Zhao, Ping Wei, and Lei Zhang. Network design principle for robust oscillatory behaviors with respect to biological noise. *Elife*, 11:e76188, 2022.

[46] Tal Danino, Octavio Mondragón-Palomino, Lev Tsimring, and Jeff Hasty. A synchronized quorum of genetic clocks. *Nature*, 463(7279):326–330, 2010.

[47] H Irisawa, HF Brown, and W Giles. Cardiac pacemaking in the sinoatrial node. *Physiological reviews*, 73(1):197–227, 1993.

[48] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.

[49] Namiko Mitarai, Kim Sneppen, and Steen Pedersen. Ribosome collisions and translation efficiency: optimization by codon usage and mrna destabilization. *Journal of molecular biology*, 382(1):236–245, 2008.

[50] Eva Maria Novoa and Lluís Ribas de Pouplana. Speeding with control: codon usage, trnas, and ribosomes. *Trends in Genetics*, 28(11):574–581, 2012.

[51] RJR Algar, T Ellis, and G-B Stan. Modelling essential interactions between synthetic genes and their chassis cell. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pages 5437–5444. IEEE, 2014.

[52] Jan-Hendrik Trösemeier, Sophia Rudorf, Holger Loessner, Benjamin Hofner, Andreas Reuter, Thomas Schulenborg, Ina Koch, Isabelle Bekeredjian-Ding, Reinhard Lipowsky, and Christel Kamp. Optimizing the dynamics of protein expression. *Scientific reports*, 9(1):1–15, 2019.

[53] Hadas Zur, Rachel Cohen-Kupiec, Sophie Vinokour, and Tamir Tuller. Algorithms for ribosome traffic engineering and their potential in improving host cells' titer and growth rate. *Scientific reports*, 10(1):21202–15, December 2020.

[54] Peter Sarvari, Duncan Ingram, and Guy-Bart Stan. A modelling framework linking resource-based stochastic translation to the optimal design of synthetic constructs. *Biology*, 10(1):37, 2021.

[55] Dieter Haas and Geneviève Défago. Biological control of soil-borne pathogens by fluorescent pseudomonads. *Nature reviews microbiology*, 3(4):307–319, 2005.

[56] Razia Alam Gilani, Mazhar Rafique, Abdul Rehman, Muhammad Farooq Hussain Munis, Shafiq Ur Rehman, and Hassan Javed Chaudhary. Biodegradation of chlorpyrifos by bacterial genus pseudomonas. *Journal of basic microbiology*, 56(2):105–119, 2016.

[57] Alicia Broto, Erika Gaspari, Samuel Miravet-Verde, Vitor Martins dos Santos, and Mark Isalan. A genetic toolkit and gene switches to limit mycoplasma growth for a synthetic vaccine chassis. *ResearchSquare*, 2021.

[58] Timothy Frei, Federica Cella, Fabiana Tedeschi, Joaquín Gutiérrez, Guy-Bart Stan, Mustafa Khammash, and Velia Siciliano. Characterization and mitigation of gene expression burden in mammalian cells. *Nature Communications*, 11(1):1–14, 2020.

[59] Eyal Metzl-Raz, Moshe Kafri, Gilad Yaakov, Ilya Soifer, Yonat Gurvich, and Naama Barkai. Principles of cellular resource allocation revealed by condition-dependent proteome profiling. *Elife*, 6, 2017.

[60] Sarah Kostinski and Shlomi Reuveni. Growth laws and invariants from ribosome biogenesis in lower eukarya. *Physical Review Research*, 3(1):013020, 2021.

[61] Matthew Scott and Terence Hwa. Bacterial growth laws and their applications. *Current opinion in biotechnology*, 22(4):559–565, 2011.