

# Component Characterization in a Growth-Dependent Physiological Context: Optimal Experimental Design

## **Authors:**

Nathan Braniff, Matthew Scott, Brian Ingalls

*Date Submitted:* 2019-04-15

*Keywords:* cell physiology, host-context effects, optimal control, optimal experimental design, characterization, model fitting, synthetic biology

## **Abstract:**

Synthetic biology design challenges have driven the use of mathematical models to characterize genetic components and to explore complex design spaces. Traditional approaches to characterization have largely ignored the effect of strain and growth conditions on the dynamics of synthetic genetic circuits, and have thus confounded intrinsic features of the circuit components with cell-level context effects. We present a model that distinguishes an activated gene's intrinsic kinetics from its physiological context. We then demonstrate an optimal experimental design approach to identify dynamic induction experiments for efficient estimation of the component's intrinsic parameters. Maximally informative experiments are chosen by formulating the design as an optimal control problem; direct multiple-shooting is used to identify the optimum. Our numerical results suggest that the intrinsic parameters of a genetic component can be more accurately estimated using optimal experimental designs, and that the choice of growth rates, sampling schedule, and input profile each play an important role. The proposed approach to coupled component?host modelling can support gene circuit design across a range of physiological conditions.

*Record Type:* Published Article

*Submitted To:* LAPSE (Living Archive for Process Systems Engineering)

*Citation (overall record, always the latest version):*

LAPSE:2019.0539

*Citation (this specific file, latest version):*

LAPSE:2019.0539-1

*Citation (this specific file, this version):*

LAPSE:2019.0539-1v1

*DOI of Published Version:* <https://doi.org/10.3390/pr7010052>

*License:* Creative Commons Attribution 4.0 International (CC BY 4.0)

Article

# Component Characterization in a Growth-Dependent Physiological Context: Optimal Experimental Design

Nathan Braniff, Matthew Scott and Brian Ingalls \* 

Department of Applied Mathematics, University of Waterloo, Waterloo, ON, N2L 3G1, Canada; nbraniff@uwaterloo.ca (N.B.); mscott@uwaterloo.ca (M.S.)

\* Correspondence: bingalls@uwaterloo.ca

Received: 28 November 2018; Accepted: 15 January 2019; Published: 21 January 2019



**Abstract:** Synthetic biology design challenges have driven the use of mathematical models to characterize genetic components and to explore complex design spaces. Traditional approaches to characterization have largely ignored the effect of strain and growth conditions on the dynamics of synthetic genetic circuits, and have thus confounded intrinsic features of the circuit components with cell-level context effects. We present a model that distinguishes an activated gene's intrinsic kinetics from its physiological context. We then demonstrate an optimal experimental design approach to identify dynamic induction experiments for efficient estimation of the component's intrinsic parameters. Maximally informative experiments are chosen by formulating the design as an optimal control problem; direct multiple-shooting is used to identify the optimum. Our numerical results suggest that the intrinsic parameters of a genetic component can be more accurately estimated using optimal experimental designs, and that the choice of growth rates, sampling schedule, and input profile each play an important role. The proposed approach to coupled component–host modelling can support gene circuit design across a range of physiological conditions.

**Keywords:** synthetic biology; model fitting; characterization; optimal experimental design; optimal control; cell physiology; host-context effects

## 1. Introduction

Proposed applications of synthetic biology demand complex synthetic constructs involving dynamic internal regulation. Novel analytic and experimental approaches will be needed to efficiently navigate the corresponding design space [1,2]. Model-based design approaches promise to (partially) replace costly experiments with computer simulations. A range of theoretical approaches to automated or computer-assisted design have been published [3–10]. When supported by reliable characterization of genetic regulatory components, automated design algorithms can greatly increase the efficiency of design. As an example, Nielsen et al. demonstrated efficient automated design of very large genetic logic circuits from a carefully designed and characterized regulatory library [11]. Systematic characterization of genetic components can include the generation of standardized data sheets [12], the use of standardized relative units [13], and the predictive characterization of circuit dynamics [14]. This type of characterization is demanding and resource-intensive; to date it has rarely been accomplished. The resulting knowledge deficit is a major bottleneck to the wider use of model-based design.

A drawback of standard approaches to characterization is that they fail to distinguish host physiology from the intrinsic properties of the construct [15]. The physiological state incorporates, among other features, (i) the DNA quantity and gene copy number, (ii) available RNA polymerases (RNAPs) and ribosomes, and (iii) the cell volume, all of which impact gene expression [16,17]. Calibration of model parameters without accounting for the cell's physiology results in aggregate

parameters that describe lumped effects from both host and component. Such a model cannot be trusted to extrapolate beyond the physiological state in which it was calibrated. Separating host state and component behaviour is a difficult and multivariate problem. Experimentalists cannot directly perturb or even measure many physiological properties. The cell's physiological state can be modulated indirectly by external or internal perturbations including modulation of (i) nutrient sources [17,18], (ii) antibiotics [18], (iii) gene expression burden [18], and (iv) metabolic fluxes [19,20], as well as temperature, pH, and osmolarity. The aggregate effect of each of these perturbations influences growth rate, but predicting host properties from the perturbation or the growth rate is not generally possible. However, for nutrient-limited growth, it has been shown that the exponential growth rate acts as a summary statistic for the physiological state of an *E. coli* host cell [16,17]. Klumpp and Hwa demonstrated how nutrient-limited growth rates can then be used as an aggregate predictor of the physiological effects on gene expression [16].

In [16], Klumpp and Hwa stop short of proposing an explicit dynamic model for coupling physiology to gene expression. They focus primarily on steady-state behaviour. More recent works have developed coupled models of gene expression, host physiology, and growth rates [21–23]. Here, we focus specifically on the use of a coarse-grained model to empirically predict physiological properties from an observed exponential growth rate. Our model distinguishes intrinsic parameters of the genetic construct from extrinsic parameters of the host's physiological state. This distinction allows for extrapolation across physiological conditions and for reuse of the estimated component parameters. We have aimed to keep the parameter set small to maximize both identifiability and interpretability.

Noise and nonlinearity make precise estimation of the parameters that characterize biomolecular systems a challenging task [24,25]. Optimal experimental design (OED) tools offer a means to improve the efficiency of data collection for model calibration [26,27]. Despite its potential to increase experimental efficiency and the precision of parameter estimates, OED has not seen widespread use in laboratory experiments within systems or synthetic biology. Two notable exceptions are reported by Bandara et al. and Ruess et al.; both groups implemented optimal experimental design for efficient calibration of dynamic biological models [28,29]. OED techniques are especially promising when coupled with optogenetic or microfluidic techniques, which allow for a broad range of dynamic perturbations in vivo [30]. Here, we employ optimal experimental design algorithms originally demonstrated for chemical and bioprocess engineering applications [31–33] for the characterization of a genetic component from simulated data.

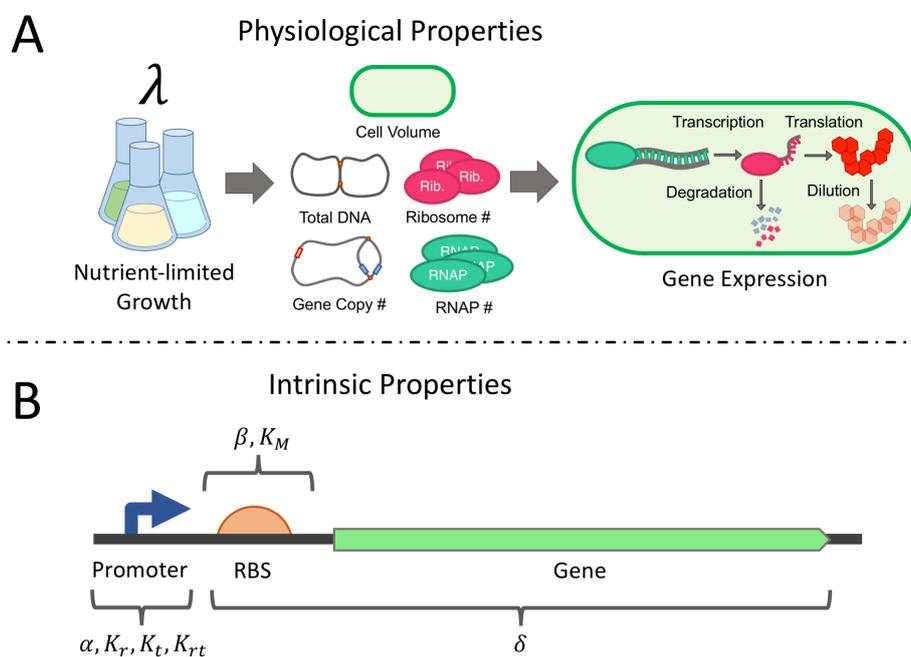
We develop our physiologically aware model of gene expression for an exponentially growing *E. coli* population, because this is the only host system for which the necessary data has been collected. We use nutrient quality in exponential phase as a predictable controller of the cell's growth rate and relevant physiological state. We use the model to optimally design dynamic induction experiments across a set of growth rates to simulate efficient estimation of the intrinsic parameters of the genetic component. We adopt an optimal design approach that expresses the experimental design as an optimal control problem, which can be efficiently solved using numerical optimal control methods, such as multiple shooting. We treat the induction profile, growth rate, and sampling schedule as experimental controls. Their selection is simultaneously optimized over multiple sub-experiments. This simultaneous optimization of sampling schedule and multiple experimental perturbations, including growth rates, is an improvement over previously published accounts of OED in synthetic biology (although see [34]). Previously published analyses either have assumed constant sampling rates or have chosen sampling schedules in a secondary optimization step [29,35], both of which are sub-optimal [36]. We use numerical simulations to demonstrate that the optimal experiments outperform intuitively designed experiments: the optimal designs improve both parameter estimation accuracy and out-of-sample prediction accuracy. We further demonstrate that the use of multiple growth rates is important for model identifiability over realistic parameter ranges.

## 2. Materials and Methods

In this work, we design optimal experiments for calibrating a dynamic model of expression of a genomically integrated gene, induced by an activating transcription factor (TF). For simplicity, we assume that the controlled induction input is the activating transcription factor copy number,  $u$ . (More realistically, the controlled input would be some signal that influences the TF abundance.) Our population-averaged model incorporates both gene-specific intrinsic parameters and parameters that characterize aspects of the host physiology. The physiological parameters of the model are dependent on the steady-state exponential growth rate,  $\lambda$ , controlled via nutrient quality. The model describes mRNA copy number,  $X_{rna}$ , and protein copy number,  $X_{prot}$ ,

$$\begin{aligned} \frac{d X_{rna}}{dt} \frac{1}{V} &= \alpha \frac{g}{V} \frac{\frac{P_a}{\eta G} K_r + \frac{P_a K_{rt}}{(\eta G)^2} u(t)}{1 + \frac{P_a}{\eta G} K_r + \left( \frac{K_t}{\eta G} + \frac{P_a K_{rt}}{(\eta G)^2} \right) u(t)} - \delta \zeta \frac{X_{rna}}{V} \\ \frac{d X_{prot}}{dt} \frac{1}{V} &= \frac{\beta \frac{R_f}{V}}{K_M + \frac{R_f}{V}} \frac{X_{rna}}{V} - \lambda \frac{X_{prot}}{V}. \end{aligned} \quad (1)$$

The gene's intrinsic characteristics are captured by the intrinsic parameters  $\alpha$ ,  $K_r$ ,  $K_t$ ,  $K_{rt}$ ,  $\delta$ ,  $\beta$ , and  $K_M$ , while expression is also dependent on the physiological parameters:  $V$ ,  $g$ ,  $P_a$ ,  $G$ ,  $R_f$ , and  $\lambda$ , which are growth-rate-dependent, and  $\eta$  and  $\zeta$ , which are fixed. The intrinsic parameters characterize each gene's induction behaviour. In contrast, the physiological parameters reflect the state of the host cell (Figure 1). These parameters are summarized in Table 1, along with nominal values and relevant ranges. Justification for these parameter values is provided in the next section.



**Figure 1.** (A) Growth media nutrient quality dictates the growth rate of an *E. coli* culture. The observed growth rate can be used to predict many physiological parameters of the host, which in turn influence gene expression. (B) By accounting for physiological parameters, we can optimally estimate intrinsic parameters of a genetic construct, which reflect properties of its sequence. These intrinsic parameters can be reused across growth conditions and can be used to guide changes to the construct sequence.

**Table 1.** Summary of Intrinsic and Physiological Parameters

Parameter Label	Intrinsic Parameter	Nominal Value	Feasible Range
<b>Intrinsic Transcription Parameters</b>			
Promoter Escape Rate	$\alpha$	20 min <sup>-1</sup>	[1–30]
RNAP-Promoter Binding	$K_r$	40	[10–40]
TF-Promoter Binding	$K_t$	$5 \times 10^5$	$[2 \times 10^3 - 1 \times 10^6]$
TF-RNAP Interaction	$K_{rt}$	$1.09 \times 10^9$	$[4.02 \times 10^5 - 5.93 \times 10^{10}]$
<b>Intrinsic mRNA Decay Parameters</b>			
mRNA Decay Rate	$\delta$	$2.57 \times 10^{-4} \mu\text{m}^{-3} \text{min}^{-1}$	$[7.7 \times 10^{-5} - 7.7 \times 10^{-4}]$
<b>Intrinsic Translation Parameters</b>			
Max. Initiation Rate	$\beta$	4.0 min <sup>-1</sup>	[1–10]
Half-saturating Constant	$K_M$	$750 \mu\text{m}^{-3}$	[750–1500]
Property Label	Physiological Property	Value at $\mu = 0.6 \text{ db/h}$	Value at $\mu = 3 \text{ db/h}$
<b>Physiological Properties of Transcription</b>			
Gene Copy Number	$g$	1.4	5.7
Available RNAP	$P_a$	1000	4000
Genome-lengths of DNA	$G$	1.3	4.3
<b>Physiological Properties of mRNA Decay</b>			
RNase Concentration	$\xi$	$900 \mu\text{m}^{-3} \text{min}^{-1}$	$900 \mu\text{m}^{-3} \text{min}^{-1}$
<b>Physiological Properties of Translation</b>			
Free Ribosomes	$R_f$	600	7000
<b>General Physiological Properties</b>			
Cell Volume	$V$	$0.4 \mu\text{m}^{-3}$	$2.24 \mu\text{m}^{-3}$
Growth Rate	$\lambda$	$0.7 \times 10^{-2} \text{min}^{-1}$	$3.5 \times 10^{-2} \text{min}^{-1}$

### 2.1. Derivation of the Physiological Gene Expression Model

We develop the model for the case of an exponentially growing *E. coli* population, where growth rate is controlled by nutrient limitation. While some features of the model may generalize to other cell types, it is only in the case of *E. coli* that sufficient data has been collected to provide reasonable estimates of functional relations and parameter values.

We employ two standard measures of growth rate. The doubling rate  $\mu$  is the inverse of the doubling time  $\tau_\mu$ . The exponential growth rate  $\lambda = \ln(2)/\tau_\mu$  is used to describe population growth as  $P_0 e^{\lambda t}$ , and is thus suitable to use as the dilution rate in a differential equation model. We treat the exponential growth rate,  $\lambda$ , as an independent variable determined by experimental conditions. We also assume the growth rate is constant throughout each experiment, which implies that no nutrient shifts occur.

#### 2.1.1. Cell Volume and Mass, DNA Content, and Protein Mass

Prior work in bacterial physiology has established expressions for the physiological parameters  $V$ ,  $G$ , and  $g$  in terms of  $\lambda$ . (While each of the parameters is expected to undergo random fluctuations and to vary with the cell cycle, these expressions approximate average values over time.)

Cell volume has been shown to scale with growth rate exponentially [37,38]:

$$V = V_0 e^{(C+D)\lambda}. \quad (2)$$

Here the constant  $V_0$  is the “initiation volume” measured to be  $V_0 = 0.28 \mu\text{m}^3$  by Si et al. [37]. The parameters  $C$  and  $D$  represent the time periods required to replicate the chromosome and to septate, respectively. We take  $C = 40 \text{ min}$  and  $D = 20 \text{ min}$  [38].

An expression for the average number of genome equivalent lengths of DNA,  $G$ , is given in terms of  $C$ ,  $D$ , and  $\lambda$  by Cooper and Helmstetter [38]:

$$G = \frac{1}{\lambda C} (e^{(C+D)\lambda} - e^{D\lambda}). \quad (3)$$

Individual loci vary in copy number based both on growth rate and on their location relative to the origin of replication. (At faster growth rates there are more copies of genes near the origin because the cell must have multiple rounds of DNA replication underway.) We use the constant  $l_{ori}$  to indicate the gene's position relative to the origin:  $l_{ori} = 0$  at the origin;  $l_{ori} = 1$  at the terminus. Bremer and Churchwood [39] provide a simple derivation for the average gene copy number,  $g$ , of a specific locus as

$$g = e^{((C+D)-l_{ori}C)\lambda}. \quad (4)$$

### 2.1.2. Total RNA Polymerase (RNAP)

We first note that the buoyant density of *E. coli* has been observed as constant across growth rates: Kubitschek et al. found  $\rho = 1.09 \text{ pg } \mu\text{m}^{-3}$  [40] while Basan et al. report  $\rho = 0.215 \text{ pg } \mu\text{m}^{-3}$  [41] (average of reported values). We use the dry weight data from [17] with Si et al.'s description of volume [37] to select an intermediate density of  $\rho = 0.55 \text{ pg } \mu\text{m}^{-3}$ .

The average cell mass for a given growth rate is then

$$M_{Tot} = \rho V = \rho V_0 e^{(C+D)\lambda}. \quad (5)$$

This total cell mass  $M_{Tot}$  can be partitioned into fractions of protein and other constituents. We fit the protein fraction of the mass, denoted  $\Phi_{pr}$ , to data from [17] with a linear function (Figure S1):

$$\Phi_{pr} = \kappa_{pr}\lambda + \Phi_{pr0} \quad (6)$$

with  $\kappa_{pr} = -6.47 \text{ min}$  and  $\Phi_{pr0} = 0.65$ .

The total RNAP fraction of the overall protein mass, which we denote  $\Phi_p$ , exhibits an approximately linear dependence on growth rate, which we fit to data from [17] (Figure S2) as

$$\Phi_p = \kappa_p\lambda + \Phi_{p0}. \quad (7)$$

with  $\kappa_p = 0.30 \text{ min}$  and  $\Phi_{p0} = 0.0074$ . We can then express the total mass of all RNAP protein,  $M_{RNAP}$ , as

$$M_{RNAP} = \underbrace{(\kappa_p\lambda + \Phi_{p0})}_{\text{RNAP \% of Prot.}} \overbrace{(\kappa_{pr}\lambda + \Phi_{pr0}) \rho V_0 e^{(C+D)\lambda}}^{\text{Protein Mass}} \underbrace{\rho V_0 e^{(C+D)\lambda}}_{\text{Cell Mass}}. \quad (8)$$

The total number of RNAPs per cell can be determined by dividing  $M_{RNAP}$  by the protein mass per RNAP core enzyme, which we denote  $m_{rnep}$  and is estimated as  $6.3 \times 10^{-7} \text{ pg}$  [42].

### 2.1.3. Available RNAP

Each RNAP can be classified by its state: freely diffusing, weakly DNA bound at a non-specific site, actively transcribing other genes, paused or non-functioning during transcription, or immature [43–45]. Below, we describe promoter binding in terms of a thermodynamic model that takes into account the observed rapid equilibrium between non-specifically bound and freely diffusing RNAPs [43]. We therefore define the combined pool of free and non-specifically bound RNAPs as the *available*

RNAP pool, with molecular population size  $P_a$ . Denoting the fraction of RNAP in this available pool by  $\Phi_a$ , we estimate, from Bakshi et al. [43] and Stracy et al. [46] (details in the Supplementary Materials Section 1.3):

$$\Phi_a = \kappa_a \lambda + \Phi_{a0} \quad (9)$$

with  $\kappa_a = -9.3 \text{ min}$  and  $\Phi_{a0} = 0.59$ . Using this relation we have an expression for the available RNAP:

$$P_a = \frac{\rho V_0}{m_{rnep}} (\kappa_a \lambda + \Phi_{a0}) (\kappa_p \lambda + \Phi_{p0}) (\kappa_{pr} \lambda + \Phi_{pr0}) e^{(C+D)\lambda}. \quad (10)$$

#### 2.1.4. Transcription Rate

We describe the initiation of transcription via a thermodynamic equilibrium model of promoter occupancy involving available RNAPs, transcription factors (TFs), promoter copies, and non-specific binding sites along the genomic DNA [47,48]. At a given growth rate, we assume there are  $P_a$  available RNAP copies and  $T_a$  active transcription factor copies diffusing along the genomic DNA, and that the DNA contains  $N_s$  non-specific binding sites to which these DNA binding proteins may weakly attach and  $g$  copies of the regulated promoter of interest. Further, we assume that  $N_s \gg P_a, T_a, g$  and that each binding of an RNAP or a TF to a non-specific site or a promoter is characterized by an associated binding energy;  $\epsilon_{rn}$  and  $\epsilon_{rg}$  for RNAP binding to the non-specific sites and promoters, respectively, and  $\epsilon_{tn}$  and  $\epsilon_{tg}$  for transcription factor binding to non-specific sites and promoters, respectively (all  $\epsilon$  are negative [49]).

We use these species and site counts to enumerate the possible arrangements of RNAP and TF across the genome, and we use the binding energies to derive Boltzmann weights for each arrangement [47]. We denote the differences between the energy involved in binding the promoter and the background non-specific binding as  $\Delta\epsilon_t = \epsilon_{tg} - \epsilon_{tn}$  and  $\Delta\epsilon_r = \epsilon_{rg} - \epsilon_{rn}$ . (Note,  $\epsilon_{tn} > \epsilon_{tg}$  and  $\epsilon_{rn} > \epsilon_{rg}$  so that  $\Delta\epsilon_t$  and  $\Delta\epsilon_r$  are both negative [49].) We denote the Boltzmann weights as  $K_r = e^{-\frac{\Delta\epsilon_r}{k_B T}}$ ,  $K_t = e^{-\frac{\Delta\epsilon_t}{k_B T}}$ , and  $K_{rt} = e^{-\frac{(\Delta\epsilon_r + \Delta\epsilon_t + \epsilon_{pt})}{k_B T}}$  (with  $k_B$ , Boltzmann's constant, and  $T$ , temperature in degrees Kelvin) where  $\epsilon_{rt}$  is the binding energy between RNA polymerase and transcription factor when both are bound to the same promoter. Then, the equilibrium probability of a single promoter being occupied by an RNAP is (further details in the Supplementary Materials Section 1.4)

$$p_{bound} = \frac{\frac{P_a}{N_s} K_r + \frac{P_a T_a}{N_s^2} K_{rt}}{1 + \frac{P_a}{N_s} K_r + \frac{T_a}{N_s} K_t + \frac{P_a T_a}{N_s^2} K_{rt}}. \quad (11)$$

With  $g$  promoter copies, and presuming that open complex and promoter escape occurs at a fixed rate  $\alpha$  [50] we have the initiation rate as

$$\text{Initiation rate} = \alpha g \frac{\frac{P_a}{N_s} K_r + \frac{P_a T_a}{N_s^2} K_{rt}}{1 + \frac{P_a}{N_s} K_r + \frac{T_a}{N_s} K_t + \frac{P_a T_a}{N_s^2} K_{rt}}. \quad (12)$$

We next establish estimates of the parameter values for transcription. To begin, multiplying  $G$  by the genomic density of non-specific binding sites,  $\eta = 5 \times 10^6$  sites/genome, yields an estimate of the total number of non-specific binding sites  $N_s$  as a function of growth rate [47].

From Heyduk et al., we have the rate of open complex formation  $\alpha$  for the phage lambda  $P_R$  promoter as  $19.2 \text{ min}^{-1}$  [50]. Assuming  $P_R$  exemplifies a strong promoter, we estimate a reasonable range based on analysis of constitutive promoters and mutants of the  $P_R$  sequences as  $1 - 30 \text{ min}^{-1}$  [50–53].

The constants  $K_r$  and  $K_t$  can be interpreted as ratios of dissociation constants for the DNA-binding species (RNAP and TF) binding to non-specific DNA versus the promoter sequence [47]. This provides a convenient method for constraining their feasible values from reported dissociation rates. Expressing

them as such yields;  $K_r = K_{rnap}^{ns} / K_{rnap}^{prom}$ ,  $K_t = K_{tf}^{ns} / K_{tf}^{prom}$  and  $K_{rt} = K_r K_t \exp(\epsilon_{rt} / k_B T)$ . Here  $K_{rnap}^{ns}$  and  $K_{tf}^{ns}$  are dissociation constants of the TF and RNAP with respect to non-specific DNA binding and  $K_{rnap}^{prom}$  and  $K_{tf}^{prom}$  are dissociation constants for promoter binding. The non-specific dissociation constant for RNAP,  $K_{rnap}^{ns}$ , has been observed to be approximately 10,000 nM [47]. The promoter-specific dissociation constant,  $K_{rnap}^{prom}$ , varies from promoter to promoter. For *lacP1*, it is approximately 550 nM and for T7 it is approximately 3 nM [47]. To represent a relatively weak constitutive leak for an inducible promoter, we expect values nearer to the *lacP1* dissociation constant would be reasonable, and so presume  $K_{rnap}^{prom}$  could range from 250 to 1000 nM. Using the above value for  $K_{rnap}^{ns}$  and the range for  $K_{rnap}^{prom}$ , we estimate a feasible range for  $K_r$  to be between 10 and 40 (unitless). We set  $K_{rnap}^{prom}$  to be 250 nM, so our nominal value for  $K_r$  is 40, suggesting a slightly stronger leak than found in uninduced *lacP1*. (Over the growth rates we consider, we found  $K_r$  to be practically unidentifiable, with high variability in its estimated value. We therefore fixed it to the nominal value for our experimental design and fitting.)

To estimate  $K_t$ , Stormo suggests a reasonable range for the promoter-specific dissociation constant,  $K_{tf}^{prom}$ , of 0.01–1000 nM [54]. We assume we are working with a relatively strong activator and that  $K_{tf}^{prom}$  lies in the range 0.01–5 nM. Stormo also suggests that non-specific binding dissociation constants are between three and six orders of magnitude less than the specific binding constants. For simplicity, we follow Bintu et al. and let  $K_{tf}^{ns} = 10,000$  nM [47]. This yields a range for  $K_t$  between 2000 and  $10^6$ . We chose  $K_{tf}^{prom} = 0.02$  and  $K_t = 5 \times 10^5$  as nominal values.

As noted, the parameter  $K_{rt}$  can be expressed as  $K_{rt} = K_r K_t \exp(-\epsilon_{rt} / k_B T)$  with  $\epsilon_{rt}$  representing the energy involved in the RNAP-TF interaction at the promoter. Few estimates exist for such values. Bintu et al. use demonstrative values of  $\epsilon_{rt} / k_B T$  ranging from  $-3.5$  to  $-4.5$ . We allow a wider range from  $-3$  to  $-5$ . This allows  $K_{rt}$  to range from  $4.02 \times 10^5$  to  $1.61 \times 10^{10}$ . We take a nominal value of  $K_{rt} = 1.09 \times 10^9$ .

### 2.1.5. mRNA Degradation

The data in [17] indicate that the mRNA decay rate is linear in mRNA copy number, with a growth-rate-independent rate constant. In [16], the authors hypothesize that this is due to the maintenance of a constant concentration of RNase E, the primary ribonuclease involved in mRNA decay initiation in *E. coli* [55]. (RNase E exhibits auto-regulation which appears to keep its concentration constant [56–58]. Moreover, RNase E appears to be expressed in excess, resulting in insensitivity to small changes in its abundance [58,59].) Under this assumption and using mass action, we have a simple model for mRNA decay as follows

$$\text{mRNA Decay Rate (copy \# per min)} = \delta \frac{E}{V} X_{rna}. \quad (13)$$

Here  $E$  is the copy number of RNase E. Assuming  $\frac{E}{V}$  is constant across growth rates (and therefore  $E \propto V$ ), we can express the mRNA degradation rate as

$$\text{mRNA Decay Rate (copy \# per min)} = \delta \zeta X_{rna}. \quad (14)$$

Here  $\zeta$  is the constant concentration of relevant degrading enzymes in the host.

The parameter  $\delta$  represents the susceptibility of the the mRNA transcript to RNase degradation, and acts as a mass action constant. The half-life of mRNAs can range from 1 – 10 min; Chen et al. suggest the mean RNA half life is near to 2.5 min [60–62]. Based on a constant concentration of RNase E of  $\zeta = 900 \mu\text{m}^{-3}$  [63] and taking a nominal half-life of 3 min, we suggest  $\delta \approx 2.57 \times 10^{-4} \mu\text{m}^{-3} \text{min}^{-1}$ .

### 2.1.6. Total and Free Ribosome Populations

A linear relation for the fraction,  $\Phi_r$ , of protein mass that is composed of ribosomal protein was derived in [18]:

$$\Phi_r = \kappa_r \lambda + \Phi_{r0}. \quad (15)$$

Fitting this model to data from Bremer [17] yields estimates of  $\kappa_r = 5.48 \text{ min}$  and  $\Phi_{r0} = 0.030$ , (see Figure S4). The total ribosomal protein mass  $M_{Rib}$  is then

$$M_{Rib} = (\kappa_r \lambda + \Phi_{r0}) (\kappa_{pr} \lambda + \Phi_{pr0}) \rho V_0 e^{(C+D)\lambda}. \quad (16)$$

Each ribosome has a mass of 2.7 MDa of which 35% is protein [64]. The individual ribosomal protein mass is therefore  $m_{rib} = 1.57 \times 10^{-6} \text{ pg}$ . We then have the number of ribosomes per cell,  $R_{Tot}$ , as

$$R_{Tot} = (\kappa_r \lambda + \Phi_{r0}) (\kappa_{pr} \lambda + \Phi_{pr0}) \frac{\rho V_0 e^{(C+D)\lambda}}{m_{rib}}. \quad (17)$$

Based on results in Dai et al. [65], we have assumed that about 10% of the ribosomes are inactive. Assuming that these are free, we denote this fraction as  $\Phi_f = 0.1$  and have

$$R_f = \Phi_f R_{Tot} = \frac{\Phi_f \rho V_0}{m_{rib}} (\kappa_r \lambda + \Phi_{r0}) (\kappa_{pr} \lambda + \Phi_{pr0}) e^{(C+D)\lambda}. \quad (18)$$

### 2.1.7. Translation Rate

We employ the Michaelis–Menten model of translation initiation proposed by Borkowski et al. [66]. In terms of the free ribosome concentration  $R_f$ ,

$$\text{Translation Rate (copy \# per min)} = \frac{\beta \frac{R_f}{V}}{K_M + \frac{R_f}{V}} X_{rna}. \quad (19)$$

In this model, the mRNA's ribosome binding site (RBS) is characterized by two constants:  $\beta$ , the maximal translation initiation rate per mRNA, and  $K_M$ , a half-saturating constant specific to the given RBS. A detailed justification for this model is provided in the Supplementary Materials Section 1.6.

To estimate  $\beta$ , we note that translation initiation rates on the order of  $\beta = 4 \text{ min}^{-1}$  have been observed for *lacA* [67]. Additional studies of RBS activity suggest a wide variation in expression levels induced by RBS alterations, ranging over orders of magnitude [68,69]. We presume that a reasonable range for the maximal translation initiation rate is from  $1 \text{ min}^{-1}$  to  $10 \text{ min}^{-1}$ .

The values of  $K_M$  reported by Borkowski et al. [66] are reported in arbitrary units. They observe a range of saturation levels from near linear behaviour to near constant saturation [66] from which we can propose a range of  $K_M$  values based on the observed free ribosome concentration, which in our model is between  $1000 \mu\text{m}^{-3}$  and  $3000 \mu\text{m}^{-3}$ . We let  $K_M$  range between  $750 \mu\text{m}^{-3}$  and  $1500 \mu\text{m}^{-3}$ , with a nominal value of  $K_M = 750 \mu\text{m}^{-3}$ , which is approaching saturation (and hence the near constant translation efficiency observed by Klumpp and Bremer [16,17]).

## 2.2. Optimal Experimental Design

To illustrate calibration of the intrinsic parameters of the model, we consider a set of dynamic induction experiments conducted over multiple growth rates, where the experimenter can select (i) the (constant) growth rate, (ii) the time-varying induction profile, and (iii) the sampling schedule. We define an experiment as a set of  $N = 3$  sub-experiments, each with a constant growth rate  $\lambda$  (possibly

repeated), organized into vector  $\lambda = [\lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}]$ . (The exponential growth rates,  $\lambda^{(i)}$ , are reported as doubling rates  $\mu^{(i)}$  (db/hr) in the results for interpretability). In each sub-experiment, the population begins at rest and responds to a dynamic induction signal in the form of a time varying transcription factor copy number,  $u(t) = [u^{(1)}(t), u^{(2)}(t), u^{(3)}(t)]$ . (Such an input  $u^{(i)}(t)$  could be implemented by, e.g., a calibrated induction system [70] or closed loop control of fluorescently tagged TFs [71].) For computational efficiency, we have restricted each  $u^{(i)}(t)$  to be a piecewise constant, with 6 constant control values delivered for 100 min each, for a total duration of  $t_f = 600$  min. We have constrained the maximum value of each  $u^{(i)}(t)$  to  $u_{max} = 1000$  and the minimum to  $u_{min} = 0$ . This wide range was selected to ensure it spans the unsaturated range of the promoter well in excess. For each sub-experiment, we also select sampling schedules for both mRNA and protein species. A single sampling schedule for a specific species is defined as a list of points,  $s_{(species)} = \{p_1, \dots, p_l, \dots, p_L\}$ . Each  $p_l = (\tau_l, c_l)$  consists of a time  $\tau_l \in [0, 600]$  and a positive integer count  $c_l$  of the number of samples taken at that time. We have restricted the design such that  $c_l \in \{1, 2, 3\}$ . The total number of samples for each species,  $c_{Tot}$  is constrained to be no more than 12. The sampling schedule for each sub-experiment,  $S^{(i)}$ , consists of a schedule for each species  $S^{(i)} = \{s_{(rna)}^{(i)}, s_{(prot)}^{(i)}\}$ . These are collected into an overall schedule:  $S = \{S^{(1)}, S^{(2)}, S^{(3)}\}$ .

Optimization over the sampling schedules as defined above involves integer programming, which poses significant computational challenges. We avoid these by employing a relaxation approach [33,72,73]. We treat the sampling schedule as a continuous sampling density such that  $s_{(species)}^{(i)}$  corresponds to a density  $w_{(species)}^{(i)}(t)$ . Each sub-experiment has sampling density  $W^{(i)}(t) = \{w_{(rna)}^{(i)}(t), w_{(prot)}^{(i)}(t)\}$  and the overall sampling schedule is  $W(t) = [W^{(1)}, W^{(2)}, W^{(3)}]$ . These sampling densities are restricted to be a piecewise constant over 48 equal intervals (each of length 12.5 min) during each sub-experiment. Each sampling density  $w_{(species)}^{(i)}(t)$  is upper bounded by  $w_{max} = \frac{3}{12.5}$ , and the integral of the sampling distribution,  $\hat{w}_n^{(species)}$  is bounded by the maximal number of samples,  $c_{Tot}$ :

$$\hat{w}_{(species)}^{(i)} = \int_0^{t_f} w_{(species)}^{(i)}(t) dt \leq c_{Tot}. \quad (20)$$

The relaxed schedule must be discretized to arrive at an implementable sampling schedule [33]. In practice, the optimal densities are often bang-bang [73] and so can be easily discretized by the sampling interval length (12.5 min) to recover integer sample counts. To account for non-integer values, we applied the Sum-Up-Rounding strategy, a common heuristic in integer programming, to recover integer solutions that approximate the optimal discrete schedule [72]. This discretization strategy effectively rounds the continuous density while respecting the sampling constraints [72]. After generating the sampling counts,  $c_l$ , from the sampling densities, we fixed their associated times  $\tau_l$  to the centre of each of the 12.5 min intervals.

To characterize a genetic construct with the induction experiments as defined above, we seek an experimental design that can accurately estimate the intrinsic parameter set  $\theta$ . As stated in Section 2.1.4, the parameter  $K_r$  proved to be practically unidentifiable under the model specifications described above. Consequently, we fixed it to its nominal value, and took the intrinsic parameter set to be estimated as  $\theta = [\alpha \beta K_t K_{rt} \delta]$ . These parameters characterize the regulated promoter, its regulating transcription factor and the downstream gene sequence, independent of growth context. We denote an estimate of the true parameters as  $\hat{\theta}$ . For our objective, we seek to minimize the determinant of the covariance matrix  $\det(\text{cov}(\hat{\theta}))$ , in what is known as a D-optimal design [74]. The determinant of the covariance matrix is also known as the generalized variance [75]; minimizing it is equivalent to minimizing the volume of the confidence ellipsoid defined by the covariance matrix [76]. Because the model is nonlinear and the true parameter vector is, by definition, unknown, it is not possible to estimate the parameter covariance matrix a priori. Instead of estimating  $\text{cov}(\hat{\theta})$  directly, we use the Fisher information computed at an initial guess  $\theta_o$  as a proxy. In linear models (or in the limit of

large sample sizes or low signal to noise ratio), the inverse of the FIM is asymptotically equivalent to  $\text{cov}(\hat{\theta})$  [77]:

$$\mathcal{I}(\theta, t_f)^{-1} \approx \text{cov}(\hat{\theta}). \quad (21)$$

In this case minimizing  $\det(\text{cov}(\hat{\theta}))$  is equivalent to maximizing the determinant of the Fisher information matrix,  $\det(\mathcal{I})$ . In the non-linear finite sample regime in which our experiments are conducted, this relation is admittedly tenuous; the use of an initial guess  $\theta_o$  also introduces potential errors. However, the process of iterating the local approximation and successively updating  $\theta_o$  after each experiment has been numerically demonstrated to yield convergence to the true parameter set in some cases [26,27]. We thus define our objective as  $\Theta_D(\mathcal{I}) = \det(\mathcal{I})$ , which is generally known as the D-optimality score [76]. (Other options such as A, E, and E-modified optimality minimize other properties of the covariance [74,76,78].)

To determine the optimal set of experiments, we follow an optimal control-based procedure described in [31–33]. (This approach to OED was demonstrated by application to chemical and bioprocess models. It has seen limited use in systems biology [28] and has not previously been applied to component characterization in a synthetic biology context.) To formulate a control problem we must state the objective, the D-optimal score  $\Theta_D(\mathcal{I}) = \det(\mathcal{I})$ , in terms of the model dynamics. We restate the model in the following generic form:

$$\frac{d\mathbf{X}}{dt} = \mathbf{F}(\mathbf{X}, \theta, \lambda^{(i)}, u^{(i)}(t)). \quad (22)$$

Here  $\mathbf{X} = [X_{rna}, X_{prot}]$  and  $\mathbf{F}$  is the right hand side of the model expressed in Equation (1).

We determine the local sensitivities of  $X_{rna}$  and  $X_{prot}$  with respect to each parameter  $\theta_i$ , by solving the following system of sensitivity equations

$$\frac{d}{dt} \frac{\partial \mathbf{X}}{\partial \theta_j} = \frac{\partial \mathbf{F}}{\partial \theta_j} + \frac{\partial \mathbf{F}}{\partial \mathbf{X}} \frac{\partial \mathbf{X}}{\partial \theta_j}. \quad (23)$$

Because the parameters are dimensioned, their scalings can vary widely, leading to poor conditioning of the Fisher information matrix and associated computations [31]. To rectify this, we scale the sensitivities by the parameter values (i.e., we use logarithmic sensitivities)

$$\bar{\mathbf{X}}_{\theta_j} = \frac{\partial \mathbf{X}}{\partial \log(\theta_j)} = \theta_j \frac{\partial \mathbf{X}}{\partial \theta_j}. \quad (24)$$

Applying this change of variables yields

$$\frac{d\bar{\mathbf{X}}_{\theta_j}}{dt} = \theta_j \frac{\partial \mathbf{F}}{\partial \theta_j} + \frac{\partial \mathbf{F}}{\partial \mathbf{X}} \bar{\mathbf{X}}_{\theta_j}. \quad (25)$$

The initial conditions for the state variables and their sensitivities are constrained to be at steady state with respect to a zero induction input ( $u = 0$ ):

$$\mathbf{X}(t = 0) = \mathbf{X}_{SS}(\theta, \lambda, u = 0) \quad (26)$$

$$\bar{\mathbf{X}}_{\theta_j}(t = 0) = \bar{\mathbf{X}}_{\theta_j, SS}(\theta, \lambda, u = 0). \quad (27)$$

Because the steady state depends on both the growth rate and the initialized parameter values, these initial conditions are not constants. They therefore appear as nonlinear constraints in the optimization problem.

We denote the sampling variance for observations of species by  $\sigma_{species}^2$ . We assume normally distributed errors with variances equal to 5% of the species quantity:

$$\sigma_{rna}^2 = (0.05)X_{rna} \quad (28)$$

$$\sigma_{prot}^2 = (0.05)X_{prot}. \quad (29)$$

We further assume no covariance between species measurements. The (scaled) Fisher information matrix  $\mathcal{I}(\boldsymbol{\theta}, t)$  can then be written for the relaxed problem as a differential equation, as per [33]

$$\frac{d}{dt}\mathcal{I}_{jk}(\boldsymbol{\theta}, t) = \bar{X}_{\theta_j}\boldsymbol{\Omega}(t)\boldsymbol{\Sigma}^{-1}(t)\bar{X}_{\theta_k} \quad \mathcal{I}_{jk}(\boldsymbol{\theta}, 0) = 0 \quad (30)$$

where the matrices  $\boldsymbol{\Omega}(t)$  and  $\boldsymbol{\Sigma}(t)$  are defined as

$$\boldsymbol{\Omega}(t) = \begin{bmatrix} w_n^{(rna)}(t) & 0 \\ 0 & w_n^{(prot)}(t) \end{bmatrix}, \quad \boldsymbol{\Sigma}(t) = \begin{bmatrix} \sigma_{rna}^2 X_{rna}(t) & 0 \\ 0 & \sigma_{prot}^2 X_{prot}(t) \end{bmatrix}. \quad (31)$$

Here  $w_n^{(species)}(t)$  are sampling densities, as defined above. Further details on the integration of the Fisher information over continuous sampling densities can be found in [33,73]. Because samples are assumed to be independent across time points as well as species, the Fisher information for the experiment is additive across both. (Here we use the Fisher information for homoskedastic error, despite the heteroskedasticity of the model. We found this simpler formulation performed equivalently and was more tractable.) Therefore, we can express the cumulative Fisher information for a given sub-experiment from time 0 to  $t_f$  as

$$\mathcal{I}_{jk}^{(i)}(\boldsymbol{\theta}, t_f) = \int_0^{t_f} \bar{X}_{\theta_i}^{(j)} \boldsymbol{\Omega}^{(i)}(t) \boldsymbol{\Sigma}^{-1}(t) \bar{X}_{\theta_k}^{(i)} dt. \quad (32)$$

The Fisher information for the total experiment,  $\mathcal{I}_{Tot}$ , is then the sum over all sub-experiments

$$\mathcal{I}_{Tot} = \sum_{i=1}^N \mathcal{I}^{(i)}(\boldsymbol{\theta}, t_f). \quad (33)$$

The overall objective (D-optimality score) is then

$$\Theta_D(\mathcal{I}_{Tot}) = \det \left( \sum_{i=1}^N \mathcal{I}^{(i)} \right). \quad (34)$$

The matrix is symmetric, so only the diagonal and lower triangular elements need to be computed.

The value of  $\Theta_D(\mathcal{I}_{Tot})$  can vary over many orders of magnitudes for different experiments, and so to improve numerical accuracy we take the logarithm of the determinant as the objective. Finally, because most optimization packages are minimizers, we invert the sign:

$$\min_{u,\lambda,W} -\ln(\Theta_D(\mathcal{I}_{Tot})). \quad (35)$$

For numerical stability, we compute the determinant using QR factorization (details in Section 2 of the Supplementary Materials).

In summary, we formulate our OED optimal control problem as

*Objective:*

$$\min_{u, \lambda, W} -\ln(\Theta_D(\mathcal{I}_{Tot})) = -\ln\left(\det\left(\sum_{i=1}^N \mathcal{I}^{(i)}(\boldsymbol{\theta}, t_f)\right)\right)$$

*Subject to* ( $\forall i \in \{1, \dots, N\}$ ):

$$\begin{aligned} \frac{d\mathbf{X}^{(i)}}{dt} &= \mathbf{F}(\mathbf{X}^{(i)}, \boldsymbol{\theta}, \lambda^{(i)}, u^{(i)}(t)) \\ \frac{d\bar{\mathbf{X}}_{\theta_j}^{(i)}}{dt} &= \theta_j \frac{\partial \mathbf{F}(\mathbf{X}^{(i)}, \boldsymbol{\theta}, \lambda^{(i)}, u^{(i)}(t))}{\partial \theta_j} + \frac{\partial \mathbf{F}(\mathbf{X}^{(i)}, \boldsymbol{\theta}, \lambda^{(i)}, u^{(i)}(t))}{\partial \mathbf{X}} \bar{\mathbf{X}}_{\theta_j}^{(i)}, \quad \forall \theta_j \in \boldsymbol{\theta} \\ \frac{d\hat{w}_{sp}^{(i)}}{dt} &= w_{sp}^{(i)}(t), \quad \forall sp \in \{rna, prot\} \\ \frac{d\mathcal{I}_{jk}^{(i)}}{dt} &= \bar{\mathbf{X}}_{\theta_j}^{(i)} \boldsymbol{\Omega}^{(i)}(t) \left(\boldsymbol{\Sigma}^{(i)}(t)\right)^{-1} \bar{\mathbf{X}}_{\theta_k}^{(i)}, \quad \forall \theta_j, \theta_k \in \boldsymbol{\theta} \end{aligned}$$

*With Constraints:*

$$\begin{aligned} \mathbf{X}^{(i)}(t=0) &= \mathbf{X}_{SS}^{(i)}(\boldsymbol{\theta}, \lambda^{(i)}, u=0) \\ \bar{\mathbf{X}}_{\theta_j}^{(i)}(t=0) &= \bar{\mathbf{X}}_{\theta_j, SS}^{(i)}(\boldsymbol{\theta}, \lambda^{(i)}, u=0), \quad \forall \theta_j \in \boldsymbol{\theta} \\ \mathcal{I}_{jk}^{(i)}(\boldsymbol{\theta}, 0) &= 0 \quad \forall \theta_j, \theta_k \in \boldsymbol{\theta} \\ \hat{w}_{sp}^{(i)}(0) &= 0, \quad sp \in \{rna, prot\} \\ \hat{w}_{sp}^{(i)}(t_f) &< c_{Tot}, \quad sp \in \{rna, prot\} \\ 0 &< w_{sp}^{(i)}(t) < w_{max}, \quad sp \in \{rna, prot\} \\ 0 &\leq u^{(i)}(t) \leq u_{max} \\ \lambda_{min} &< \lambda^{(i)} < \lambda_{max}. \end{aligned} \tag{36}$$

This problem formulation matches that used by Telen et al. and Hoang et al. [31,33]. We used a multiple shooting algorithm to solve the optimal control problem [32]. We implemented this algorithm in CasADi, a rapid prototyping optimal control toolbox, using its MATLAB interface [79]. CasADi uses algorithmic differentiation to compute first and second derivatives with respect to both the objective and the constraints. This higher-order information is used for optimization in an interior-point barrier method implemented in the nonlinear programming package IPOPT [80]. Further details of the multiple shooting algorithm and optimization settings can be found in Section 2 of the Supplementary Materials.

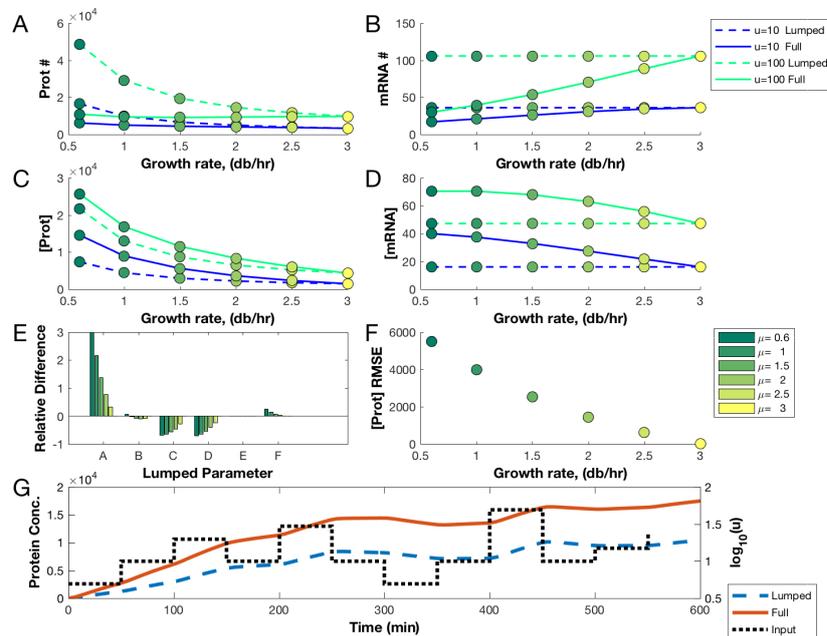
### 3. Results

#### 3.1. Comparing Lumped and Physiologically Aware Models

Models of gene expression typically use lumped parameters that confound physiological effects with intrinsic features. Such models are usually calibrated against data collected from cells grown in a particular medium (and so at a single growth rate). To illustrate the consequences of neglecting growth effects, we consider a lumped version of our growth dependent model (Equation (1))

$$\begin{aligned}\frac{d[X_{rna}]}{dt} &= A \frac{B + Cu}{1 + B + (D + C)u} - E[X_{rna}] \\ \frac{d[X_{prot}]}{dt} &= F[X_{rna}] - \lambda[X_{prot}]\end{aligned}\quad (37)$$

where parameters  $A = \frac{\alpha g}{V}$ ,  $B = \frac{P_s K_r}{\eta G}$ ,  $C = \frac{P_a K_{rt}}{(\eta G)^2}$ ,  $D = \frac{K_t}{\eta G}$ ,  $E = \delta \zeta$ , and  $F = \frac{\beta R_f}{V(K_M V + R_f)}$  are treated as growth-rate-independent constants. (The one exception is the protein decay rate, which is equal to the exponential growth rate.) Of course, this parameter lumping does not impact model behaviour for the growth rate at which the model is calibrated, but accurate extrapolation to other growth rates cannot be assured. This loss of accuracy is illustrated in Figure 2. Panels A and B show predictions of the full growth-dependent model and the lumped model, both calibrated to fast growth rates (3 db/h). As the growth rate drops, significant deviation in the predicted steady state copy number of both protein and mRNA is observed. In Figures 2C,D, we see similar deviation for the steady state concentrations of these species. (We note that the relative deviations in concentration are comparatively small. Protein concentration is important for, e.g., metabolic enzymes. In contrast, copy number is more important for DNA binding proteins, which tend to disperse along the DNA rather than the total cell volume [43,71].) A complementary measure of inaccuracy due to lumping appears in Figure 2E, which shows how the lumped parameter values vary with growth rate when the full model is used to determine their values (in comparison to their values calibrated at the reference fast growth rate of 3 db/h). The divergence in these parameters sets has a significant effect on the model's dynamic behaviour. Figure 2F shows the root mean squared error (RMSE) between the two models' response to the input profile depicted in Figure 2G. The difference in dynamic behaviour between the lumped and full model under this dynamic induction, at a slow growth rate of 0.5 db/h, is shown in Figure 2G.

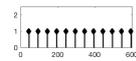
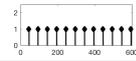
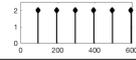
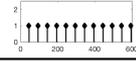


**Figure 2.** (A,B) Steady state protein (A) and mRNA (B) copy number predictions for the lumped and full model across growth rate for two different constant input levels. (C,D) Corresponding protein (C) and mRNA (D) concentrations. (E) Relative difference in the lumped parameter values, fit at 3 db/h, compared with the values predicted by the full growth dependent model. (F) Root mean squared protein concentration error over the dynamic simulation (G) between the full and lumped model. (G) Predicted protein concentrations of the lumped and full model over a dynamic simulation with a growth rate of 0.5 db/h.

### 3.2. Null and Optimal Experimental Designs

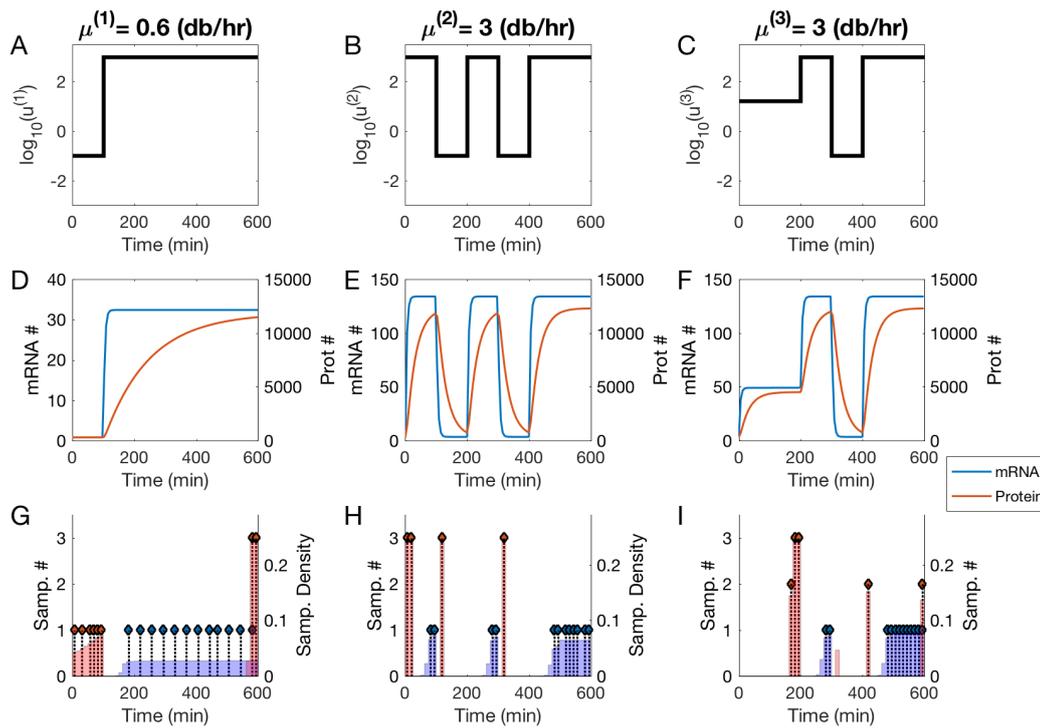
We next explore how experimental design can improve the estimation of the intrinsic parameters of the full growth-dependent model (Equation (1)). We have selected a null experiment, shown in Table 2, to represent a reasonable non-optimized design. It consists of a set of logarithmically (base 10) distributed pulses, delivered at evenly spaced growth rates, with evenly spaced samples taken every half-hour (mRNA and protein samples are taken at the same time). We propose this as a sensible first experiment for fitting a dynamic model. We designed the null experiments assuming limited information on the dynamics of the specific gene expression system. They were selected to involve a reasonably comprehensive set of perturbations and measurements. For comparison, we examine three variants of the null experiment, also shown in Table 2, each with a perturbation to either the growth rates, induction profile, or sampling rate of the null. The growth variant is identical to the null case, except it is performed over a narrower range of growth rates. The sampling variant differs from the null by redistributing the samples so that their rate is halved but the sampling number at each point is doubled (with the same total number of samples). The induction variant uses linearly (rather than logarithmically) spaced strengths of induction pulses. Shown with each design is its predicted optimality score (the negative log determinant of the scaled Fisher information matrix computed at the true parameter set: Equation (35)). As expected, each variant provides less information than the null (as measured by D-optimality).

**Table 2.** Null (non-optimal) experimental designs.

Experiment	Growth Rates (db/h)	Induction Pattern ( $\log_{10}$ )	Sampling Schedule	Optimality
Null Experiment	{0.6, 1.8, 3}			−63.8
Growth Variant	{2, 2.5, 3}			−61.5
Sampling Variant	{0.6, 1.8, 3}			−62.6
Induction Variant	{0.6,1.8,3}			−60.5

In contrast, Figure 3 depicts an optimal experiment for our model. This design was generated by using the true nominal parameter vector as the algorithm's initialization parameters. This is an artificial scenario (the true parameters are not known initially, by definition), but it serves to demonstrate the difference between intuitive designs and optimal designs selected by our method. In Figure 3, each column describes a sub-experiment, labeled with the corresponding optimal growth rate. The top row, (A–C), depicts the optimal input profiles; the middle row, (D–F), shows the system response in both mRNA and protein copy number. The last row, (G–I), shows the sampling densities (in the shaded regions) as well as the results of the Sum-Up-Rounding discretization scheme (depicted by the stem plots, where the height is an integer representing the sample count). The optimality score of this design is −69.6. (In this case, the optimal design selected the maximum growth rate twice. We suspect this is due to higher sensitivities at faster growth. More complex models typically demand observations over a wider range of growth rates.)

Comparing the null and optimal designs we see that, while the null experiments have an intuitive appeal because of their wide, even distribution of experimental choices, none achieves a score similar to the optimized design. Further, while null variants exhibit differences in optimality scores, when compared to the optimized experiment these differences are small. This suggests that manually tuning aggregate design measures is less effective than the holistic optimization provided by the OED approach.

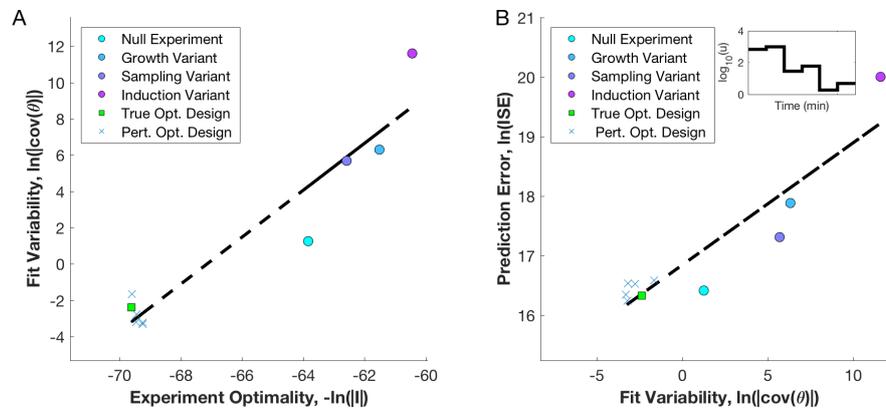


**Figure 3.** An optimal experiment (designed at the true value of  $\theta$ ). Each column depicts one sub-experiment, labelled with optimal doubling rates:  $\mu^{(1)}$ ,  $\mu^{(2)}$  and  $\mu^{(3)}$  (corresponding to exponential growth rates  $\lambda^{(1)}$ ,  $\lambda^{(2)}$  and  $\lambda^{(3)}$ ). The top row (A–C) depicts the optimal inputs,  $u^{(1)}(t)$ ,  $u^{(2)}(t)$  and  $u^{(3)}(t)$  (on a  $\log_{10}$  scale). The middle row (D–F) shows the system response (both mRNA and protein copy number) for each induction profile. The last row shows the sampling densities  $w_{rna}^{(i)}$  and  $w_{prot}^{(i)}$  (in shaded areas, blue and red, respectively), for both protein and mRNA, as well as the rounded sampling schedule,  $s_{rna}^{(i)}$  and  $s_{prot}^{(i)}$  (depicted by the stem plot). The optimality score of this experiment was  $-69.6$ .

### 3.3. Utility of Optimal Designs for Parameter Identification and Prediction

The difference in optimality scores between the null and optimized experiments suggests significant improvements in parameter estimation using the optimized design. However, the statistical theory used to derive the optimality score can only be guaranteed to hold a priori for linear models in the limit of small observation variances and large samples [77]. To validate that our optimality scores correspond to improved parameter estimation accuracy, we used simulated experiments to assess the correlation between theoretical and observed variability. To generate simulated data, we simulated the model using the nominal true parameters and added normally distributed observation error with a variance equal to 5% of the corresponding species count. We used a multiple shooting approach for parameter estimation; obvious outliers were removed (details in Section 2 of the Supplementary Materials). Figure 4A shows, on the vertical axis, the logarithm of the generalized variance (determinant of the observed covariance matrix) for the collection of parameter estimates (each corresponding to an experimental design). This measure of fit variability is computed from the collection of parameter fits to independent simulations of the given experiment. From this set of estimates we computed the observed covariance matrix and the generalized variance. The optimality score of the design is shown on the horizontal axis (lower is better), computed at the true parameter value. Figure 4A shows that the optimality score (objective of the OED approach) and the observed parameter covariance (sampled measure of design ‘quality’) correlate well. In particular, we note that the optimal design achieves both a better optimal score and a smaller generalized parameter variance when compared to the non-optimal designs. This suggests that the objective function (using the homeostatic FIM) is

a useful measure of design quality, despite the nonlinearity and heteroskedasticity of the model. We also note that adopting a uniform sampling schedule or reducing the range of growth rates is expected to result in considerably worse performance, as evidenced by comparison between null variants.



**Figure 4.** (A) Results from the true optimal (optimal experiment designed at true parameter value) and null experiments (circles): optimality score on the horizontal axis; observed generalized variance of the parameter estimate on the horizontal axis. (B) The same experiments: generalized parameter variance on the horizontal axis; log integral of the squared error on an out-of-sample prediction experiment (inset) on the vertical axis. Optimal experiments designed with erroneous initial parameter guesses are shown as X's. (Linear trend lines also shown.)

In addition to accuracy of parameter estimates, accurate predictions of the system behaviour for out-of-sample conditions is also important for model-based design. For linear regression models, D-optimal designs that minimize the generalized variance of the parameter estimate are equivalent to designs that minimize the upper bound on prediction variance (General Equivalence Theorem) [81]. This guarantee does not generalize to our dynamic, nonlinear, and heteroskedastic model. To verify if prediction accuracy indeed correlates with D-optimality for our model, we ran a second simulation study. For each of the parameter estimates used in calculating the fit variability for the experimental designs, we simulated the model response for a new dynamic experiment. We chose out-of-sample conditions: growth rates and induction levels not used in fitting. We simulated the model with the true nominal parameter values, and computed the integrated squared error (ISE) between the true and fitted values in each case (thus including error at all time points). Figure 4B shows the generalized parameter variance along the horizontal and the log ISE on the out-of-sample data set on the vertical axis. The plots show that the generalized parameter variance correlates reasonably well with prediction accuracy: the optimal design performs better than any of the null designs. This reflects the linear case, in which the D-optimal design sets an upper-bound on the expected prediction variance [81].

So far, we have compared designs in the idealized (and artificial) scenario in which OED is applied to the model parametrized by its true values. We relaxed that assumption by first generating five intrinsic parameter sets from a uniform distribution spanning the intervals specified in the feasible ranges from Table 1. We then ran the OED algorithm using models parametrized by these 'perturbed' parameter sets. The optimality scores of these designed experiments, when evaluated against simulated data generated by the true parameter set, are plotted in Figures 4A,B, depicted as X's.

#### 4. Discussion

We have proposed a physiologically aware model of gene expression in *E. coli* that accounts for the effects of nutrient limitation on the host physiology. The model is more complex than standard models of gene expression, but this comes with several benefits. The model naturally suggests a partitioning of the parameter set into (1) intrinsic parameters that characterize the genetic component and (2) a set of empirically derived parameters characterizing the host's physiological state as a function of

nutrient-mediated growth rate. The intrinsic parameters can thus be reused across a range of growth conditions. In particular, because the intrinsic parameters can be linked to sequence properties of the component (e.g., promoter affinities, RBS strength, and mRNA stability), they could provide insight as to which aspect of a component's DNA sequence could be altered to achieve a desired effect across a variety of contexts. Future work could attempt to link such intrinsic parameters to sequence properties directly. In contrast, the extrinsic parameters can be used to predict the host's context based on the observed growth rate in a range of nutrient conditions.

Accurate estimation of the intrinsic parameter set is more difficult than estimation in context-naive models. We have demonstrated that optimal experimental design can mitigate this challenge by identifying maximally informative experiments. We applied a comprehensive OED platform, optimizing over constant growth rate, time-dependent induction profile, and sampling schedule. Many current experimental design approaches in systems biology use general purpose optimization algorithms that scale poorly to such multivariate and non-linear designs [30,74]. This often leads to non-optimal selection of certain design variables [36]. To achieve computational efficiency in our optimization tasks, we re-interpreted the problem via optimal control, with growth rates, sampling schedules, and induction levels as control variables. This optimal control framing of OED problems has so far been underutilized in systems biology; it allows access to the rich tool-set developed by control and process engineers, including direct multiple-shooting and collocation methods [31–33]. These methods are ideally suited for use with emerging experimental tools such as optogenetic induction systems and automated culture and microfluidic devices [30]. Our numerical results suggest that the multiple shooting algorithm provides an efficient method to optimize experiments, improving both parameter estimation accuracy and prediction accuracy over unoptimized designs.

The optimal experimental design algorithm presented here could be improved in a number of ways. We used the simplest and most tractable form of the Fisher information. More accurate formulations account for the sample variance's parametric dependence [29] and for the constraints imposed by initial conditions [82]. Additionally, our approach provides a design that may be only locally optimal. Current iterative designs are based on the assumption that the iterated algorithm does not become trapped in some local minimum. Past numerical studies of iterative applications of OED have shown consistent convergence to the true parameter set (global optimum), but there is scope for more rigorous investigation [26,27]. Improved computational efficiency may allow users to address multiple scenarios and larger models, or even implement online experimental design algorithms that can update the design in real-time [83,84]. Our algorithm was implemented with CasADi, which provides rapid-prototyping capabilities for control problems, algorithmic differentiation and interfaces to powerful non-linear programming solvers like IPOPT [79,80]. This tool facilitates rapid implementation, but requires some mathematical expertise. Further packaging of OED tools for use by experimentalists would no doubt increase adoption in systems and synthetic biology.

The experiments proposed in this work are demanding, requiring sampling of multiple species and control of inputs over extended time periods. In practice, it is not currently possible to precisely control the transcription factor count *in vivo*. It would suffice to map the experimental input (e.g., light, chemical inducer) to the expected average copy number of TF per cell. This could be implemented through precise calibration experiments, achievable with current techniques [70]. Future experiments may be able to implement closed-loop control using real-time measurements to ensure the robust tracking of desired inputs. It also may be possible to modify the OED algorithm to account for the effects of input variability (an errors-in model). In this work, we have assumed time-series measurements of both mRNA and protein quantities. Emerging experimental tools, combined with assays like RNA-seq [85], will likely enable these complex dynamic experiments in the future [30].

For the model considered here, observing only one species results in a structurally unidentifiable parameter set. Any parameter lumping to alleviate unidentifiability leads to a loss of modularity in characterization. Even with observations of both species, the parameter  $K_r$ , which characterizes promoter leak, had to be excluded from the analysis because it was practically unidentifiable over

reasonable parameter ranges. In practice, it may be possible to ignore promoter leakiness in many cases. However, for strong leaks, it is important to identify the parameter  $K_r$ ; fixing it, as we have, may introduce significant bias in other parameter estimates. In addition to these specific identifiability concerns, the re-usability of parameter estimates is limited by the use of relative units, which are specific to the measurement instrument they are calibrated on. The methods used in this work should ideally be implemented using absolute units, which will allow for comparison between parameter values calibrated on different instruments and in different labs [86–88].

Our description of physiological state was restricted to a single case: exponentially growing *E. coli* host cells, with growth rate determined by nutrient quality. Our model formulation was made possible by significant experimental efforts into characterizing precisely this physiological response [17]. Beyond nutrient limitation, several other relevant growth perturbations are of interest to synthetic biologists, including translation-inhibiting antibiotics, gene expression burden, and metabolic knockdowns. Phenomenological growth laws and coarse-grained proteome models have been extended to some of these conditions [18,19,41]. Recent results suggest that certain metabolic perturbations may effect proteome partitioning (and possibly RNAP and ribosomal fractions) in a manner similar to nutrient limitation [19,20]. Certain physiological properties, such as the ribosome-to-protein ratio, can also be predictably linked to the growth rate controlled by expression burden or antibiotic dosage [18,41]. The linear ribosome-to-protein ratios may even generalize across species (Figure S1 of [18]). The details necessary to link these coarse-grained models to the specifics of gene expression, as originally proposed in Klumpp and Hwa [16] (and further elaborated on here) are still lacking. While the nutrient-limited growth theories have taken decades to build, modern techniques can likely accelerate this host characterization process, allowing generalization across strains, growth inhibition conditions, and potentially even microbial species.

Extensions of phenomenological growth theories to heterologous protein expression burden and modification of metabolic fluxes could have significant impact on metabolic engineering. These effects are especially relevant to the emerging sub-discipline of dynamic metabolic engineering, where synthetic gene expression circuits dynamically modulate both fluxes and heterologous enzyme expression [89–91]. Design of such systems is challenging; the engineered regulatory components modulate expression burden and enzymatic activity, which in turn affects the growth rate, broader cell physiology, and the regulatory component behaviour itself [92–94]. An extended growth theory combined with the optimal experimental design algorithm and gene expression models presented here could be valuable for guiding future work in this area.

We have proposed the coupling of physiologically aware modelling and OED techniques to address limitations in the accuracy and generalizability of component characterization. Current gene expression models often fail to account for the host or environmental state, and are calibrated with poorly constrained parameter estimates using ad hoc experimental designs. These shortcomings contribute to the limited use of model-based design in synthetic biology. Poor parameter estimates result in lackluster predictive accuracy, and context-naïve model predictions cannot generalize, resulting in recalibration for each new design and circumstance, and thus little advantage of model-based design over trial-and-error approaches. Our model has yet to be validated in vivo, but wider use of such coarse-grained, context-aware models and optimal experimental designs will ideally maximize researchers' return on experimental investment and aid efforts to rationally design gene expression circuits.

**Supplementary Materials:** Further details on the model development and OED implementation are available at <http://www.mdpi.com/2227-9717/7/1/52/s1>.

**Author Contributions:** Conceptualization: N.B. and B.I.; methodology: N.B.; software: N.B.; validation: N.B., M.S., and B.I.; writing—original draft preparation: N.B.; writing—review and editing: N.B., M.S., and B.I.

**Funding:** This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC), grant number RGPIN-2018-03826.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Appleton, E.; Densmore, D.; Madsen, C.; Roehner, N. Needs and opportunities in bio-design automation: four areas for focus. *Curr Opin. Chem. Biol.* **2017**, *40*, 111–118. [[CrossRef](#)] [[PubMed](#)]
2. Beal, J. Bridging the gap: A roadmap to breaking the biological design barrier. *Front. Bioeng. Biotechnol.* **2015**, *2*, 87. [[CrossRef](#)] [[PubMed](#)]
3. Guiziou, S.; Ulliana, F.; Moreau, V.; Leclere, M.; Bonnet, J. An Automated Design Framework for Multicellular Recombinase Logic. *ACS Synth. Biol.* **2018**, *7*, 1406–1412. [[CrossRef](#)] [[PubMed](#)]
4. Otero-Muras, I.; Henriques, D.; Banga, J.R. SYNBADm: A tool for optimization-based automated design of synthetic gene circuits. *Bioinformatics* **2016**, *32*, 3360–3362. [[CrossRef](#)] [[PubMed](#)]
5. Madec, M.; Pecheux, F.; Gendraul, Y.; Rosati, E.; Lallement, C.; Haiech, J. GeNeDA: An Open-Source Workflow for Design Automation of Gene Regulatory Networks Inspired from Microelectronics. *J. Comput. Biol.* **2016**, *23*, 841–855. [[CrossRef](#)] [[PubMed](#)]
6. Huynh, L.; Tagkopoulos, I. Fast and accurate circuit design automation through hierarchical model switching. *ACS Synth. Biol.* **2015**, *4*, 890–897. [[CrossRef](#)] [[PubMed](#)]
7. Rodrigo, G.; Jaramillo, A. AutoBioCAD: Full biodesign automation of genetic circuits. *ACS Synth. Biol.* **2012**, *2*, 230–236. [[CrossRef](#)] [[PubMed](#)]
8. Yaman, F.; Bhatia, S.; Adler, A.; Densmore, D.; Beal, J. Automated selection of synthetic biology parts for genetic regulatory networks. *ACS Synth. Biol.* **2012**, *1*, 332–344. [[CrossRef](#)] [[PubMed](#)]
9. Beal, J.; Weiss, R.; Densmore, D.; Adler, A.; Appleton, E.; Babb, J.; Bhatia, S.; Davidsohn, N.; Haddock, T.; Loyall, J.; et al. An end-to-end workflow for engineering of biological networks from high-level specifications. *ACS Synth. Biol.* **2012**, *1*, 317–331. [[CrossRef](#)]
10. Beal, J.; Lu, T.; Weiss, R. Automatic compilation from high-level biologically-oriented programming language to genetic regulatory networks. *PLoS ONE* **2011**, *6*, e22490. [[CrossRef](#)]
11. Nielsen, A.A.; Der, B.S.; Shin, J.; Vaidyanathan, P.; Paralanov, V.; Strychalski, E.A.; Ross, D.; Densmore, D.; Voigt, C.A. Genetic circuit design automation. *Science* **2016**, *352*, aac7341. [[CrossRef](#)] [[PubMed](#)]
12. Canton, B.; Labno, A.; Endy, D. Refinement and standardization of synthetic biological parts and devices. *Nat. Biotechnol.* **2008**, *26*, 787. [[CrossRef](#)] [[PubMed](#)]
13. Kelly, J.R.; Rubin, A.J.; Davis, J.H.; Ajo-Franklin, C.M.; Cumbers, J.; Czar, M.J.; de Mora, K.; Gliberman, A.L.; Monie, D.D.; Endy, D. Measuring the activity of BioBrick promoters using an in vivo reference standard. *J. Biol. Eng.* **2009**, *3*, 4. [[CrossRef](#)] [[PubMed](#)]
14. Davidsohn, N.; Beal, J.; Kiani, S.; Adler, A.; Yaman, F.; Li, Y.; Xie, Z.; Weiss, R. Accurate predictions of genetic circuit behavior from part characterization and modular composition. *ACS Synth. Biol.* **2014**, *4*, 673–681. [[CrossRef](#)] [[PubMed](#)]
15. Cardinale, S.; Arkin, A.P. Contextualizing context for synthetic biology—identifying causes of failure of synthetic biological systems. *Biotechnol. J.* **2012**, *7*, 856–866. [[CrossRef](#)]
16. Klumpp, S.; Zhang, Z.; Hwa, T. Growth rate-dependent global effects on gene expression in bacteria. *Cell* **2009**, *139*, 1366–1375. [[CrossRef](#)]
17. Bremer, H.; Dennis, P. Modulation of Chemical Composition and Other Parameters of the Cell at Different Exponential Growth Rates. *EcoSal Plus* **2008**, *3*. [[CrossRef](#)]
18. Scott, M.; Gunderson, C.W.; Mateescu, E.M.; Zhang, Z.; Hwa, T. Interdependence of cell growth and gene expression: Origins and consequences. *Science* **2010**, *330*, 1099–1102. [[CrossRef](#)]
19. You, C.; Okano, H.; Hui, S.; Zhang, Z.; Kim, M.; Gunderson, C.W.; Wang, Y.P.; Lenz, P.; Yan, D.; Hwa, T. Coordination of bacterial proteome with metabolism by cyclic AMP signalling. *Nature* **2013**, *500*, 301. [[CrossRef](#)]
20. Hui, S.; Silverman, J.M.; Chen, S.S.; Erickson, D.W.; Basan, M.; Wang, J.; Hwa, T.; Williamson, J.R. Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. *Mol. Syst. Biol.* **2015**, *11*, 784. [[CrossRef](#)]
21. Weiße, A.Y.; Oyarzún, D.A.; Danos, V.; Swain, P.S. Mechanistic links between cellular trade-offs, gene expression, and growth. *Proc. Natl. Acad. Sci. USA* **2015**. [[CrossRef](#)] [[PubMed](#)]
22. Carrera, J.; Rodrigo, G.; Singh, V.; Kirov, B.; Jaramillo, A. Empirical model and in vivo characterization of the bacterial response to synthetic gene expression show that ribosome allocation limits growth rate. *Biotechnol. J.* **2011**, *6*, 773–783. [[CrossRef](#)] [[PubMed](#)]

23. Liao, C.; Blanchard, A.E.; Lu, T. An integrative circuit–host modelling framework for predicting synthetic gene network behaviours. *Nat. Microbiol.* **2017**, *2*, 1658. [[CrossRef](#)] [[PubMed](#)]
24. Gutenkunst, R.N.; Waterfall, J.J.; Casey, F.P.; Brown, K.S.; Myers, C.R.; Sethna, J.P. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* **2007**, *3*, e189. [[CrossRef](#)]
25. Erguler, K.; Stumpf, M.P. Practical limits for reverse engineering of dynamical systems: A statistical analysis of sensitivity and parameter inferability in systems biology models. *Mol. BioSyst.* **2011**, *7*, 1593–1602. [[CrossRef](#)]
26. Hagen, D.R.; White, J.K.; Tidor, B. Convergence in parameters and predictions using computational experimental design. *Interface Focus* **2013**, *3*, 20130008. [[CrossRef](#)]
27. Apgar, J.F.; Witmer, D.K.; White, F.M.; Tidor, B. Sloppy models, parameter uncertainty, and the role of experimental design. *Mol. BioSyst.* **2010**, *6*, 1890–1900. [[CrossRef](#)]
28. Bandara, S.; Schlöder, J.P.; Eils, R.; Bock, H.G.; Meyer, T. Optimal experimental design for parameter estimation of a cell signaling model. *PLoS Comput. Biol.* **2009**, *5*, e1000558. [[CrossRef](#)]
29. Ruess, J.; Parise, F.; Miliias-Argeitis, A.; Khammash, M.; Lygeros, J. Iterative experiment design guides the characterization of a light-inducible gene expression circuit. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 8148–8153. [[CrossRef](#)]
30. Braniff, N.; Ingalls, B. New Opportunities for Optimal Design of Dynamic Experiments in Systems and Synthetic Biology. *Curr. Opin. Syst. Biol.* **2018**, *9*, 42–48. [[CrossRef](#)]
31. Hoang, M.; Barz, T.; Merchan, V.; Biegler, L.; Arellano-Garcia, H. Simultaneous solution approach to model-based experimental design. *AIChE J.* **2013**, *59*, 4169–4183. [[CrossRef](#)]
32. Janka, D.; Körkel, S.; Bock, H.G. Direct multiple shooting for nonlinear optimum experimental design. In *Multiple Shooting and Time Domain Decomposition Methods*; Springer: Berlin, Germany, 2015; pp. 115–141.
33. Telen, D.; Vercammen, D.; Logist, F.; Van Impe, J. Robustifying optimal experiment design for nonlinear, dynamic (bio) chemical systems. *Comput. Chem. Eng.* **2014**, *71*, 415–425. [[CrossRef](#)]
34. Balsa-Canto, E.; Alonso, A.A.; Banga, J.R. Computational procedures for optimal experimental design in biological systems. *IET Syst. Biol.* **2008**, *2*, 163–172. [[CrossRef](#)] [[PubMed](#)]
35. Kutalik, Z.; Cho, K.H.; Wolkenhauer, O. Optimal sampling time selection for parameter estimation in dynamic pathway modeling. *Biosystems* **2004**, *75*, 43–55. [[CrossRef](#)] [[PubMed](#)]
36. Braniff, N.; Reed, M.; Ingalls, B. Optimal experimental design for characterizing gene expression: Sample scheduling. *IFAC-PapersOnLine* **2018**, *51*, 48–51. [[CrossRef](#)]
37. Si, F.; Li, D.; Cox, S.E.; Sauls, J.T.; Azizi, O.; Sou, C.; Schwartz, A.B.; Erickstad, M.J.; Jun, Y.; Li, X.; et al. Invariance of initiation mass and predictability of cell size in *Escherichia coli*. *Curr. Biol.* **2017**, *27*, 1278–1287. [[CrossRef](#)] [[PubMed](#)]
38. Cooper, S.; Helmstetter, C.E. Chromosome replication and the division cycle of *Escherichia coli* Br. *J. Mol. Biol.* **1968**, *31*, 519–540. [[CrossRef](#)]
39. Bremer, H.; Churchward, G. An examination of the Cooper-Helmstetter theory of DNA replication in bacteria and its underlying assumptions. *J. Theor. Biol.* **1977**, *69*, 645–654. [[CrossRef](#)]
40. Kubitschek, H.E.; Baldwin, W.W.; Schroeter, S.J.; Graetzer, R. Independence of buoyant cell density and growth rate in *Escherichia coli*. *J. Bacteriol.* **1984**, *158*, 296–299.
41. Basan, M.; Zhu, M.; Dai, X.; Warren, M.; Sévin, D.; Wang, Y.P.; Hwa, T. Inflating bacterial cells by increased protein synthesis. *Mol. Syst. Biol.* **2015**, *11*, 836. [[CrossRef](#)]
42. Finn, R.D.; Orlova, E.V.; Gowen, B.; Buck, M.; van Heel, M. *Escherichia coli* RNA polymerase core and holoenzyme structures. *EMBO J.* **2000**, *19*, 6833–6844. [[CrossRef](#)] [[PubMed](#)]
43. Bakshi, S.; Dalrymple, R.M.; Li, W.; Choi, H.; Weisshaar, J.C. Partitioning of RNA polymerase activity in live *Escherichia coli* from analysis of single-molecule diffusive trajectories. *Biophys. J.* **2013**, *105*, 2676–2686. [[CrossRef](#)] [[PubMed](#)]
44. Klumpp, S.; Hwa, T. Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 20245–20250. [[CrossRef](#)] [[PubMed](#)]
45. Patrick, M.; Dennis, P.P.; Ehrenberg, M.; Bremer, H. Free RNA polymerase in *Escherichia coli*. *Biochimie* **2015**, *119*, 80–91. [[CrossRef](#)] [[PubMed](#)]

46. Stracy, M.; Lesterlin, C.; De Leon, F.G.; Uphoff, S.; Zawadzki, P.; Kapanidis, A.N. Live-cell superresolution microscopy reveals the organization of RNA polymerase in the bacterial nucleoid. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E4390–E4399. [[CrossRef](#)] [[PubMed](#)]
47. Bintu, L.; Buchler, N.E.; Garcia, H.G.; Gerland, U.; Hwa, T.; Kondev, J.; Phillips, R. Transcriptional regulation by the numbers: Models. *Curr. Opin. Genet. Dev.* **2005**, *15*, 116–124. [[CrossRef](#)] [[PubMed](#)]
48. Rydenfelt, M.; Cox III, R.S.; Garcia, H.; Phillips, R. Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration. *Phys. Rev. E* **2014**, *89*, 012702. [[CrossRef](#)]
49. Phillips, R. Napoleon is in equilibrium. *Annu. Rev. Condens. Matter Phys.* **2015**, *6*, 85–111. [[CrossRef](#)]
50. Heyduk, E.; Kuznedelov, K.; Severinov, K.; Heyduk, T. A consensus adenine at position–11 of the nontemplate strand of bacterial promoter is important for nucleation of promoter melting. *J. Biol. Chem.* **2006**, *281*, 12362–12369. [[CrossRef](#)]
51. Brunner, M.; Bujard, H. Promoter recognition and promoter strength in the *Escherichia coli* system. *EMBO J.* **1987**, *6*, 3139–3144. [[CrossRef](#)]
52. Djordjevic, M.; Bundschuh, R. Formation of the open complex by bacterial RNA polymerase—A quantitative model. *Biophys. J.* **2008**, *94*, 4233–4248. [[CrossRef](#)] [[PubMed](#)]
53. Djordjevic, M. Efficient transcription initiation in bacteria: An interplay of protein–DNA interaction parameters. *Integr. Biol.* **2013**, *5*, 796–806. [[CrossRef](#)] [[PubMed](#)]
54. Stormo, G.D. *Introduction to Protein-DNA Interactions: Structure, Thermodynamics, and Bioinformatics*; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2013.
55. Kushner, S. Messenger RNA Decay. *EcoSal Plus* **2007**. [[CrossRef](#)] [[PubMed](#)]
56. Jain, C.; Belasco, J.G. RNase E autoregulates its synthesis by controlling the degradation rate of its own mRNA in *Escherichia coli*: Unusual sensitivity of the rne transcript to RNase E activity. *Genes Dev.* **1995**, *9*, 84–96. [[CrossRef](#)] [[PubMed](#)]
57. Mudd, E.A.; Higgins, C.F. *Escherichia coli* endoribonuclease RNase E: Autoregulation of expression and site-specific cleavage of mRNA. *Mol. Microbiol.* **1993**, *9*, 557–568. [[CrossRef](#)] [[PubMed](#)]
58. Jain, C.; Deana, A.; Belasco, J.G. Consequences of RNase E scarcity in *Escherichia coli*. *Mol. Microbiol.* **2002**, *43*, 1053–1064. [[CrossRef](#)] [[PubMed](#)]
59. Ow, M.C.; Liu, Q.; Mohanty, B.K.; Andrew, M.E.; Maples, V.F.; Kushner, S.R. RNase E levels in *Escherichia coli* are controlled by a complex regulatory system that involves transcription of the rne gene from three promoters. *Mol. Microbiol.* **2002**, *43*, 159–171. [[CrossRef](#)]
60. Chen, H.; Shiroguchi, K.; Ge, H.; Xie, X.S. Genome-wide study of mRNA degradation and transcript elongation in *Escherichia coli*. *Mol. Syst. Biol.* **2015**, *11*, 781. [[CrossRef](#)]
61. Pedersen, S.; Reeh, S.; Friesen, J.D. Functional mRNA half lives in *E. coli*. *Mol. Gen. Genet. MGG* **1978**, *166*, 329–336.
62. Selinger, D.W.; Saxena, R.M.; Cheung, K.J.; Church, G.M.; Rosenow, C. Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Res.* **2003**, *13*, 216–223. [[CrossRef](#)] [[PubMed](#)]
63. Mackie, G.A. RNase E: At the interface of bacterial RNA processing and decay. *Nat. Rev. Microbiol.* **2013**, *11*, 45. [[CrossRef](#)] [[PubMed](#)]
64. Berg, J.M.; Tymoczko, J.L.; Stryer, L. *Biochemistry*, 5 ed.; WH Freeman: New York, NY, USA, 2002.
65. Dai, X.; Zhu, M.; Warren, M.; Balakrishnan, R.; Patsalo, V.; Okano, H.; Williamson, J.R.; Fredrick, K.; Wang, Y.P.; Hwa, T. Reduction of translating ribosomes enables *Escherichia coli* to maintain elongation rates during slow growth. *Nat. Microbiol.* **2017**, *2*, 16231. [[CrossRef](#)]
66. Borkowski, O.; Goelzer, A.; Schaffer, M.; Calabre, M.; Mäder, U.; Aymerich, S.; Jules, M.; Fromion, V. Translation elicits a growth rate-dependent, genome-wide, differential protein production in *Bacillus subtilis*. *Mol. Syst. Biol.* **2016**, *12*, 870. [[CrossRef](#)] [[PubMed](#)]
67. Kennell, D.; Riezman, H. Transcription and translation initiation frequencies of the *Escherichia coli* lac operon. *J. Mol. Biol.* **1977**, *114*, 1–21. [[CrossRef](#)]
68. Seo, S.W.; Yang, J.S.; Kim, I.; Yang, J.; Min, B.E.; Kim, S.; Jung, G.Y. Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency. *Metab. Eng.* **2013**, *15*, 67–74. [[CrossRef](#)] [[PubMed](#)]
69. Salis, H.M.; Mirsky, E.A.; Voigt, C.A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* **2009**, *27*, 946. [[CrossRef](#)]

70. Olson, E.J.; Hartsough, L.A.; Landry, B.P.; Shroff, R.; Tabor, J.J. Characterizing bacterial gene circuit dynamics with optically programmed gene expression signals. *Nat. Methods* **2014**, *11*, 449–455. [[CrossRef](#)]
71. de Leon, F.G.; Sellars, L.; Stracy, M.; Busby, S.J.; Kapanidis, A.N. Tracking low-copy transcription factors in living bacteria: The case of the lac repressor. *Biophys. J.* **2017**, *112*, 1316–1327. [[CrossRef](#)]
72. Sager, S.; Bock, H.G.; Diehl, M. The integer approximation error in mixed-integer optimal control. *Math. Progr.* **2012**, *133*, 1–23. [[CrossRef](#)]
73. Sager, S. Sampling decisions in optimum experimental design in the light of Pontryagin’s maximum principle. *SIAM J. Control Optim.* **2013**, *51*, 3181–3207. [[CrossRef](#)]
74. Chakrabarty, A.; Buzzard, G.T.; Rundell, A.E. Model-based design of experiments for cellular processes. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2013**, *5*, 181–203. [[CrossRef](#)] [[PubMed](#)]
75. Wilks, S.S. Certain generalizations in the analysis of variance. *Biometrika* **1932**, pp. 471–494. [[CrossRef](#)]
76. Kreutz, C.; Timmer, J. Systems biology: Experimental design. *FEBS J.* **2009**, *276*, 923–942. [[CrossRef](#)] [[PubMed](#)]
77. Vallisneri, M. Use and abuse of the Fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects. *Phys. Rev. D* **2008**, *77*, 042001. [[CrossRef](#)]
78. Franceschini, G.; Macchietto, S. Model-based design of experiments for parameter precision: State of the art. *Chem. Eng. Sci.* **2008**, *63*, 4846–4872. [[CrossRef](#)]
79. Andersson, J.A.E.; Gillis, J.; Horn, G.; Rawlings, J.B.; Diehl, M. CasADi—A software framework for nonlinear optimization and optimal control. *Math. Prog. Comput.* **2018**, in press. [[CrossRef](#)]
80. Wächter, A.; Biegler, L.T. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Prog.* **2006**, *106*, 25–57. [[CrossRef](#)]
81. Kiefer, J.; Wolfowitz, J. The equivalence of two extremum problems. *Can. J. Math.* **1960**, *12*, 234. [[CrossRef](#)]
82. Bauer, I.; Bock, H.G.; Körkel, S.; Schlöder, J.P. Numerical methods for optimum experimental design in DAE systems. *J. Comput. Appl. Math.* **2000**, *120*, 1–25. [[CrossRef](#)]
83. Galvanin, F.; Barolo, M.; Bezzo, F. Online model-based redesign of experiments for parameter estimation in dynamic systems. *Ind. Eng. Chem. Res.* **2009**, *48*, 4415–4427. [[CrossRef](#)]
84. Bandiera, L.; Hou, Z.; Kothamachu, V.; Balsa-Canto, E.; Swain, P.; Menolascina, F. On-Line Optimal Input Design Increases the Efficiency and Accuracy of the Modelling of an Inducible Synthetic Promoter. *Processes* **2018**, *6*, 148. [[CrossRef](#)]
85. Gorochoowski, T.E.; Borujeni, A.E.; Park, Y.; Nielsen, A.A.; Zhang, J.; Der, B.S.; Gordon, D.B.; Voigt, C.A. Genetic circuit characterization and debugging using RNA-seq. *Mol. Syst. Biol.* **2017**, *13*, 952. [[CrossRef](#)] [[PubMed](#)]
86. Castillo-Hair, S.M.; Sexton, J.T.; Landry, B.P.; Olson, E.J.; Igoshin, O.A.; Tabor, J.J. FlowCal: A user-friendly, open source software tool for automatically converting flow cytometry data from arbitrary to calibrated units. *ACS Synth. Biol.* **2016**, *5*, 774–780. [[CrossRef](#)] [[PubMed](#)]
87. Beal, J.; Haddock-Angelli, T.; Gershater, M.; De Mora, K.; Lizarazo, M.; Hollenhorst, J.; Rettberg, R.; et al. Reproducibility of fluorescent expression from engineered biological constructs in *E. coli*. *PLoS ONE* **2016**, *11*, e0150182.
88. Beal, J.; Haddock-Angelli, T.; Baldwin, G.; Gershater, M.; Dwijayanti, A.; Storch, M.; de Mora, K.; Lizarazo, M.; Rettberg, R.; et al. Quantification of bacterial fluorescence using independent calibrants. *PLoS ONE* **2018**, *13*, e0199432. [[CrossRef](#)] [[PubMed](#)]
89. Venayak, N.; Anesiadis, N.; Cluett, W.R.; Mahadevan, R. Engineering metabolism through dynamic control. *Curr. Opin. Biotechnol.* **2015**, *34*, 142–152. [[CrossRef](#)] [[PubMed](#)]
90. Liu, D.; Mannan, A.A.; Han, Y.; Oyarzún, D.A.; Zhang, F. Dynamic metabolic control: Towards precision engineering of metabolism. *J. Ind. Microbiol. Biotechnol.* **2018**, *56*, 535–543. [[CrossRef](#)]
91. Tan, S.Z.; Prather, K.L. Dynamic pathway regulation: Recent advances and methods of construction. *Curr. Opin. Chem. Biol.* **2017**, *41*, 28–35. [[CrossRef](#)]
92. Doong, S.J.; Gupta, A.; Prather, K.L. Layered dynamic regulation for improving metabolic pathway productivity in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 2964–2969. [[CrossRef](#)]

93. Gupta, A.; Reizman, I.M.B.; Reisch, C.R.; Prather, K.L. Dynamic regulation of metabolic flux in engineered bacteria using a pathway-independent quorum-sensing circuit. *Nat. Biotechnol.* **2017**, *35*, 273. [[CrossRef](#)]
94. Soma, Y.; Hanai, T. Self-induced metabolic state switching by a tunable cell density sensor for microbial isopropanol production. *Metab. Eng.* **2015**, *30*, 7–15. [[CrossRef](#)] [[PubMed](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).