bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Scalable dynamic characterization of synthetic gene circuits

Neil Dalchau[1], Paul K. Grant[1], Prashant Vaidyanathan[1], Colin Gravill[1], Carlo Spaccasassi[1], and Andrew Phillips[1]

[1]Biological Computation group, Microsoft Research, Cambridge, CB1 2FB, UK

Abstract

The dynamic behavior of synthetic gene circuits plays a key role in ensuring their correct function. However, our ability to accurately predict this dynamic behavior is limited by our quantitative understanding of the circuits being constructed. This understanding can be represented as a mathematical model, which can be used to optimize circuit performance and inform the design of future circuits. Previous inference methods have used fluorescent reporters to quantify average behaviors over an extended time window, resulting in a static characterization which is a poor predictor of dynamics. Here we present a method for characterizing the dynamic behavior of synthetic gene circuits. The method relies on parameter inference techniques applied to time-series measurements of cell cultures growing in microtiter plates. We use our method to design and characterize gene circuits in E. coli that provide core functionality for engineering cell behavior at the population level. We arrange 23 biological parts into 9 devices and combine them to construct and measure 9 gene circuits including relays, receivers and a degrader. We demonstrate that the behaviors of simple devices can be modeled dynamically and used to predict the behaviors of more complex circuits. Furthermore, our method allows incremental inference of models as new circuits are constructed, and lays the foundation for iteratively learning dynamic models from data in a scalable manner.

Keywords: Synthetic Biology | Gene circuits | Computational Modeling | Markov chain Monte Carlo | Quorum sensing

Correspondence: ***ndalchau@microsoft.com, aphillip@microsoft.com***

Introduction

As increasingly complex gene circuits are designed and implemented in living cells (1), the precise dynamic behavior of these circuits is becoming increasingly important for ensuring their correct function. For instance, genetically engineered microbes must produce specific quantities of proteins at particular times to optimize the yield of compounds such as medicines (2), materials (3) and chemicals (4), and genetically engineered T-cells must express the correct levels of chimeric antigen receptors at the appropriate time to avoid overstimulating the immune system and potentially killing the patient (5).

However, our ability to accurately predict the dynamic behavior of gene circuits is limited by our quantitative understanding of their properties. This understanding can be represented as a mathematical model, which can be used to optimize circuit performance and inform the design of future circuits (6, 7). The kinetic parameters of the model are required to accurately predict its dynamic behavior, and can include binding affinities, rates of transcription and translation, and rates of enzyme-catalyzed reactions. Obtaining model parameters presents a major challenge, since in most cases these parameters cannot be derived from static circuit properties such as genetic sequence.

In addition, the dynamic behavior of a circuit, such as the concentrations of RNA and proteins over time, is often not directly observable inside a living cell. As a result, model parameters need to be inferred by measuring observable outputs such as fluorescence and absorbance over time. Previous methods have used fluorescent reporters to obtain a static measurement of gene expression for a given a promoter (8–10). However, these methods focus on quantifying average behaviors over an extended time window, resulting in a static characterization which is a poor predictor of dynamics.

This paper presents a method for characterizing the dynamic behavior of synthetic gene circuits. The method relies on parameter inference techniques applied to time-series measurements of cell cultures growing in microtiter plates. By comparing fluorescent protein measurements over time to the dynamic behavior of a mathematical model, we infer the values of model parameters that cannot be measured directly. Our method allows computational models to be associated with gene circuits in a modular fashion, by reusing device-level properties based on hypotheses that can be specified in code. We arrange 23 biological parts into 9 devices and combine them to construct and measure 9 gene circuits that provide core functionality for programming cell behavior at the population level. The devices include Receivers capable of sensing and responding to two homoserine lactone (HSL) signaling molecules simultaneously, Relays capable of receiving one signal and producing the other signal in response, and a Degrader that degrades both signals. We demonstrate that the behaviors of these devices can be modeled dynamically and that device models can be composed to predict the behaviors of more complex gene circuits. Finally, our method allows incremental inference of models as new circuits are constructed and measured. We demonstrate this by iteratively constructing a repository of models and re-using param-

bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

eter values of existing devices as new devices are introduced. Propagating parameter values in this way removes the need to rerun the inference problem for an entire database of devices each time a new device is added. Our work lays the foundation for a scalable approach to learning dynamic models from data over time, by iteratively building on existing knowledge.

Results

Dynamic characterization method. We first present our dynamic characterization method, using the terminology of Fig. 1A. We define a genetic *device* as a sequence of genetic *parts*, which can include promoters, ribosome binding sites, protein coding regions and terminators. Devices are assembled into *circuits*, which are transformed into cell *cultures*. The dynamic behavior of each culture is measured under different *treatments*, such as different concentrations of input signals. Measurements of the culture can include culture density and the intensity of spectrally distinct fluorescent proteins. The method uses culture-level measurements in combination with computational (dynamical systems) models of intracellular concentrations to infer quantitative properties of parts and devices. In doing so, it enables model predictions of intracellular dynamics by reconstructing bulk culture measurements (Fig. 1B).

Inference of model parameters is complicated by the feedback that exists between a synthetic genetic circuit and the host cell in which it operates (Fig. 1C). The circuit places a burden on the host cell resources, which in turn affects circuit behavior (11–16). In order to reduce the bias that can result from inaccurately representing the burden on host cell resources (K in Fig. 1C), our method does not model these resources explicitly, as attempted in (17). Instead, we model the circuit in *open loop*, by allowing parameters quantifying cell growth (ρ) and gene expression (ξ) to take on *culture-specific* values. Culture-specific cell growth incorporates resource burden indirectly, since we observe slower growth as the synthetic circuit increases its burden on the host cell. Culture-specific gene expression captures differences in transcription, translation and degradation that arise from differences in culture conditions. This might include an increased burden on the ribosome pool leading to lower translation (16, 18), increased burden on polymerases leading to lower transcription (19), or differences in the degradation of different mRNA sequences (20). Accordingly, these can include both intrinsic and extrinsic properties (10) of synthetic gene circuits. By modelling cell growth and gene expression with culture-specific values, we can arbitrarily account for the interdependence between the two (21), despite not explicitly modeling host cell resources.

In order to estimate culture-specific gene expression parameters, we developed a dynamic version of *ratiometric characterization* (10, 22). The ratiometric method seeks to separate intrinsic properties of a circuit, such as the ability of a promoter to recruit polymerase, from extrinsic properties such as cellular metabolic burden and ribosome availability. This can be achieved by using a *standard* promoter to drive expression of a spectrally distinct fluorescent protein, which is assumed to capture the same extrinsic factors that influence fluorescent protein expression driven by other promoters in the circuit.

Dynamic characterization relies on three distinct phases (Fig. 1C), where each culture is indexed by the circuit i that it contains and the treatment j under which it is measured:

1. A *growth phase* characterizes the dynamics of culture growth, by comparing simulations of a *growth model* with measurements of culture density over time. The growth model is of the form

$$\dot{c} = \gamma(c, \rho_{ij})c \tag{1}$$

where $c(t)$ is the culture density, ρ_{ij} are the culture-specific growth parameters and the function $\gamma(c(t), \rho_{ij})$ describes the specific growth rate of the culture.

2. A *control phase* characterizes the gene expression of the culture, by comparing simulations of a *control model* with the expression of a control signal over time, such as the intensity of a fluorescent protein driven by a constitutive promoter, or the expression of a housekeeping gene. Although the culture may be genetically engineered to produce the control signal, we consider this to form part of the culture itself and to be distinct from the genetic circuits that are subsequently introduced, since the same control signal is used for all circuits and treatments. The control model is of the form

$$\begin{aligned} \dot{x}_C &= f_C(x_C, c, u_{ij}, \phi, \xi_{ij}, h_C, \rho_{ij}, \gamma) \\ y_C &= g_C(x_C) \end{aligned} \tag{2}$$

where $x_C(t)$ is a vector of unobserved state variables, $y_C(t)$ is the observable control signal, u_{ij} is the culture-specific treatment, $c(t)$ is the culture density, ϕ are culture-independent control parameters and ξ_{ij} are culture-specific control parameters that characterize the *gene expression capacity* of the culture, which is described by a function $h_C(c(t), \xi_{ij}, \phi)$. The culture-specific growth parameters ρ_{ij} are incorporated into the model via the specific growth rate $\gamma(c(t), \rho_{ij})$, which is used to model the effect of dilution due to growth.

3. A *target phase* characterizes the dynamics of the genetic circuit by comparing simulations of a *target model* with measurements of observable circuit outputs over time. The target model is of the form

$$\begin{aligned} \dot{x}_T &= f_T(x_T, c, u_{ij}, \theta, \xi_{ij}, h_T, \rho_{ij}, \gamma) \\ y_T &= g_T(x_T) \end{aligned} \tag{3}$$

bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

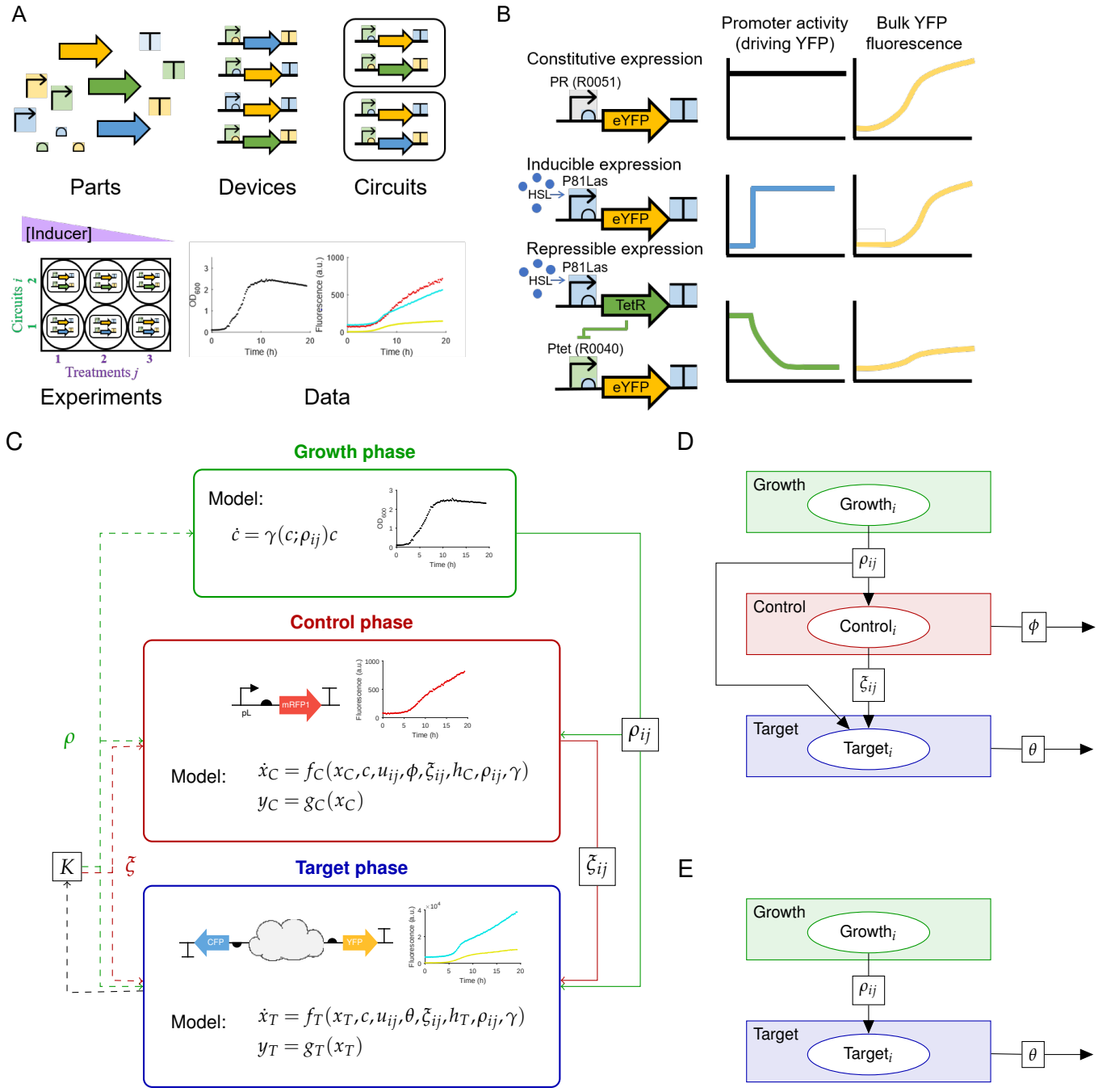

Fig. 1. Understanding the properties of genetic circuits via dynamic bulk fluorescence measurements. A. A collection of genetic parts can be composed combinatorially into a set of devices that can be introduced singly or in combination into a cellular context to produce genetic circuits. The circuits (indexed by i) can be exposed to a variety of treatment conditions (indexed by j), in a set of experiments that produce data such as optical density (OD_{600}) and bulk fluorescence at three different wavelengths corresponding to mRFP1, eCFP and eYFP . This hierarchical organization allows for the encoding of hypotheses about modularity and the reuse of measured properties in new contexts to allow for rational forward engineering. **B.** Dynamic bulk fluorescence is a rich data source but its relationship to the underlying promoter activity is non-trivial. Constitutive expression by constant promoter activity leads to a sigmoidal bulk fluorescence curve due to sigmoidal growth of cell density. Inducible expression leads to a (near-)step change in promoter activity, but is almost indistinguishable from constitutive expression. An inverter system drastically alters bulk fluorescence dynamics, which presents a challenge for time-window selection of static characterization. **C.** Our method seeks to infer the quantitative properties of a synthetic circuit while accounting for the feedback of the circuit consuming cellular resources and inhibiting growth. To approximate this *closed loop* system, an *open loop* approximation is proposed that allows each cell culture, expressing a given circuit i under a given treatment j, to be described by models of cell growth (γ) and gene expression capacity (h_C, h_T) that are parametrized by culture-specific parameters u_{ij}, ρ_{ij} and ξ_{ij}. The models of the control signal (f_C) and the target circuit (f_T) therefore embed these circuit-specific factors. **D.** Inference graph representing the three stages of ratiometric dynamic characterization for a single circuit i. The Growth node infers the growth parameters ρ_{ij} for the circuit i under a range of treatments j. This is achieved by fitting a growth model to corresponding growth data, such as optical density of the culture, for each treatment. The Control node uses the growth parameters ρ_{ij} to infer the control parameters ξ_{ij} for the same circuit i under the same treatments j. This is achieved by fitting a control model to corresponding control data, such as the concentration of fluorescent protein expressed by a constitutive promoter, for each treatment. Parameters ϕ shared between control models, such as the degradation of the constitutively expressed fluorescent protein, are also inferred. Finally, the Target node uses the growth and control parameters to infer parameters θ shared between target models, such as the strength of promoters used in the circuits. **E.** Inference graph representing the two stages of direct dynamic characterization. The graph is similar to (E) except that the Control node is omitted.

bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where $x_T(t)$ is a vector of unobserved state variables, $y_T(t)$ is a vector of observable genetic circuit outputs and θ are culture-independent circuit parameters. Similar to the control phase, the culture-specific growth parameters ρ_{ij} are incorporated into the model via the specific growth rate $\gamma(c(t),\rho_{ij})$, while the culture-specific control parameters ξ_{ij} are incorporated by scaling the expression of all genes by a function $h_T(c(t),\xi_{ij},\theta)$.

Importantly, phases (1-3) represent three different measurements of the *same culture* containing a given circuit i under a given treatment j. The three phases are collectively referred to as *ratiometric* characterization. We also consider a variant of the method in which only the growth and target phases are used, referred to as *direct* characterization, which does not attempt to quantify between-culture variations in gene expression capacity. In this case h_T is set to 1 throughout.

We developed the notion of an *inference graph* (Supplementary Information Section S1.1) to formally represent the three phases of our method. An inference graph is defined as a *labelled directed acyclic site graph*, meaning that each edge has a label and a direction, the graph does not contain any cycles, and each node contains one or more *sites*:

- Each *node* in the inference graph represents a parameter inference over multiple cultures, indexed by their circuit i and treatment j, where each circuit is associated with a computational model and corresponding experimental measurements. Parameter inference seeks to minimize the deviation between the measurements and the model simulations (see Methods). The models in a given node share *global parameters* and typically represent a collection of similar circuits, such as circuits that differ only in their ribosome binding sites at specific locations.

- Each *site* in a node represents a circuit i, where the same model is assumed to be used consistently across all treatments j for that circuit. The site can also contain *local parameters* specific to that circuit.

- Each *edge* is labelled with a set of parameters that are *propagated* from the source to the target node, where each parameter is associated with a *propagation property* that denotes how the parameter is to be propagated. For example, if the property is *Fixed* then only the parameter value is propagated. If the property is *Normal* then the posterior marginal distribution of the parameter is fitted to a Normal distribution, which is then used as a prior for the parameter in the target node. Edges can occur between nodes or sites, to represent the propagation of global or local parameters, respectively.

The inference graph representing the three phases of ratiometric dynamic characterization is shown in Fig. 1D. The Growth node fits a growth model to measurements of culture density, resulting in growth parameters ρ_{ij} for each circuit i under treatment j. These parameters are then propagated to the corresponding sites in the Control and Target nodes. The Control node then fits a control model to measurements of the control signal, resulting in control parameters ξ_{ij} for the same circuit i under the same treatment j, together with parameters ϕ shared between the control models. The control parameters are then propagated to the corresponding sites in the Target node, which fits a target model to measurements of the target signals, resulting in parameters θ shared between the target models. The inference graph representing the two phases of direct dynamic characterization is shown in Fig. 1E. More generally, inference graphs also allow parameter inference to be applied sequentially to circuits of increasing complexity, by reusing parameters shared between circuits in a consistent manner.

Modeling gene circuits. Our dynamic characterization method is general in that it can support a range of modeling hypotheses. In this paper we focus on models of gene circuits derived from elementary chemical reactions, by making several assumptions about the timescales of different biological processes, culminating in only proteins and small molecules being explicitly described by ordinary differential equations (see Methods and Section S2).

Genetic devices are conveniently composable during circuit construction and model specification. However, the quantitative properties of parts and devices may not be. The rate of translation of mRNA to protein depends on the upstream ribosome binding site, while the rate of protein synthesis also depends on other factors from its surrounding context, such as the choice of upstream promoter, downstream terminator, position on the plasmid with respect to the origin of replication or other gene expression cassettes. Accordingly, one must hypothesize the rules governing whether a parameter can be shared when parts or devices are reused in other circuits. Here, we take the view that the rate of protein synthesis, which we modeled by treating transcription and translation as a single-step process (see Methods and Supplementary Information Section S2), depends on the associated promoter, ribosome binding site and protein coding region. Only when these three elements are the same are protein synthesis parameters shared. In contrast, we hypothesize that protein degradation is the same for a given protein in different genetic circuits. For inducible promoters, we assume that the regulation function parameters only depend on promoter sequence (and the concentration of the upstream regulator), and do not depend on downstream ribosome binding site and protein coding region elements.

bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

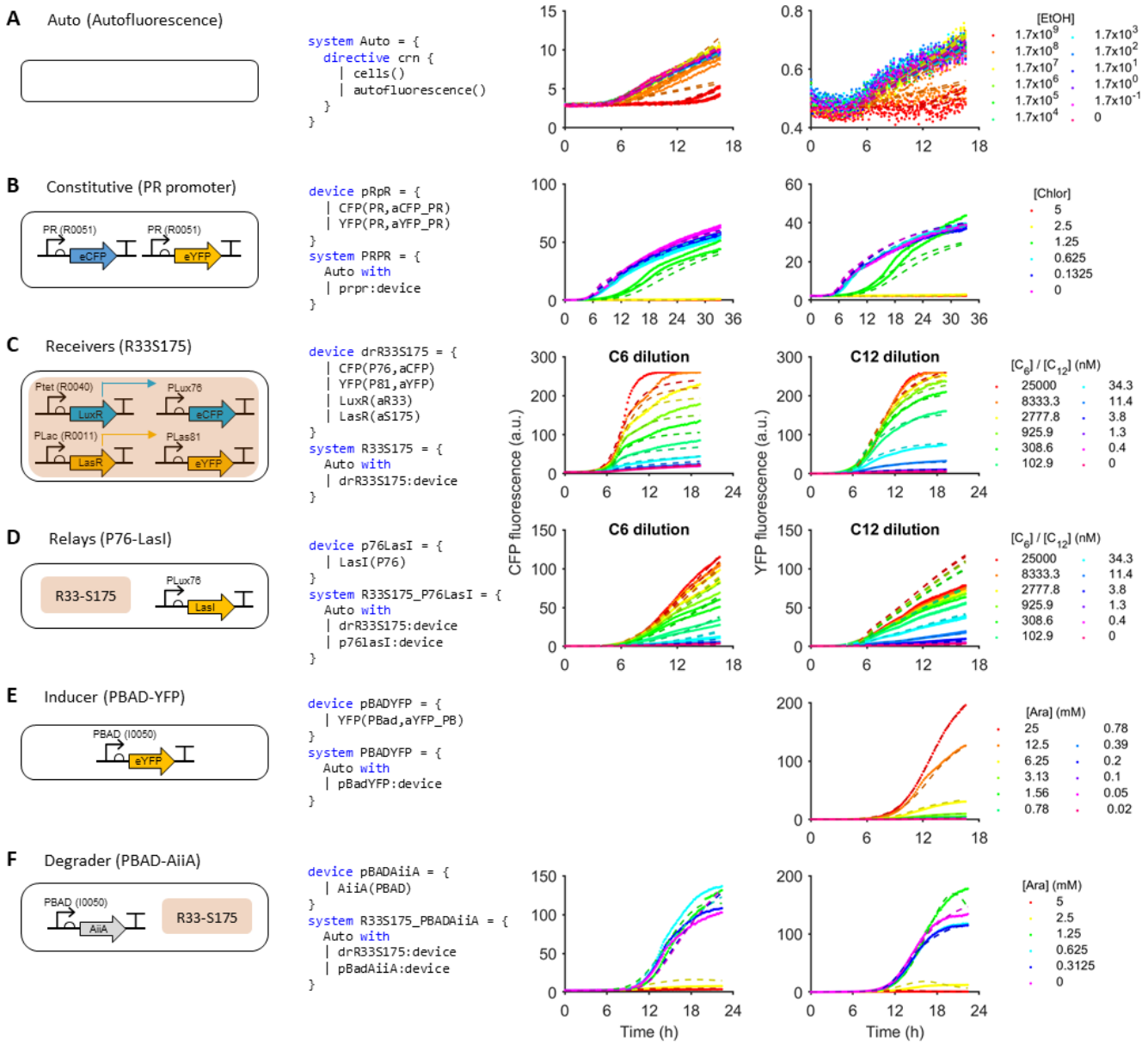

Fig. 2. Composable design and characterization of HSL signaling parts and devices. Synthetic gene circuits are constructed and modeled in a modular way. In each row, the left column graphically depicts different circuits, the middle column shows device and circuit definitions in GEC programs for each circuit and the right column shows example measurements (solid lines) and calibrated model simulations (dashed lines). **A.** Autofluorescence characterized with cells expressing no eCFP or eYFP , but measured at 480 nm and 530 nm in response to varying concentrations of EtOH. The model of the Auto circuit specifies the cellular context through module calls `cells()` and `autofluorescence()`. **B.** Fluorescent protein degradation was characterized by expressing eCFP (solid blue square) and eYFP (solid yellow square) with a constitutive promoter, treating with chloramphenicol, and measuring for a longer duration. The model of the PRPR device involves module calls `CFP()` and `YFP()`, which specify protein production with an expression PR, in addition to protein degradation and dilution (see Supplementary Information Section S1 for more details). The model of cell growth and autofluorescence is assumed to be equivalent to the model in A, and so is propagated to the system definition using the `with` keyword. **C.** HSL Receivers were characterized by measuring 4 variant circuits with different ribosome binding sites for LuxR and LasR expression (orange symbols), and treated with C6 or C12. Shown here is the R33-S175 circuit variant only. The model incorporates autofluorescence and fluorescent protein degradation from A and B. **D.** HSL Relays were characterized by inducibly expressing LasI in response to different concentrations of C6 via PLux81 (shown here) or LuxI in response to C12 via PLas81 (see Supplementary Information). The model incorporates elements of A, B and C. Notably, this system extends the `Auto` system using the `with` keyword, by combining the `drR33S175` and `p76lasI` devices. **E.** The PBAD promoter was characterized by inducing expression of eYFP with different concentrations of arabinose. The model incorporates elements of A and B. **F.** The HSL Degrader, PBAD-AiiA, was characterized by inducing its expression via PBAD, then measuring the response of the double receiver to different concentrations of HSL. Data-model comparisons are shown for the eCFP response to 25 μM C6 (left panel) and the eYFP response to 25 μM C12 (right panel) only.

Dynamic characterization case study. We demonstrated how our dynamic characterization method can be used to quantify the properties of genetic devices, by building and characterizing genetic circuits in *E. coli* that provide core functionality for engineering cell behavior at the population level. Our circuits make use of homoserine lactone (HSL) signaling components, including the Lux system from *V. fischeri* (23) and the Las system from *P. aeruginosa* (24), which have both been studied extensively in their natural contexts. They

bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

have also been used in synthetic biology contexts, either alone (25), in co-culture (26), or integrated into the same host organism (27).

In order to characterize HSL signaling devices, we designed a collection of synthetic gene circuits of increasing complexity, reusing parts in equivalent genetic contexts as much as possible (Fig. 2). For each circuit, a model was proposed that explicitly described the circuit behavior. These are summarized in the Methods section, but detailed derivations are shown in Supplementary Information Section S2. In Fig. 2, code snippets show how devices can be programmatically specified as calls to modules that describe their behaviors. Circuits are defined by specifying which devices are included and by assigning models for cell culture growth and autofluorescence. Example comparisons of calibrated model simulations against measurements are shown in Fig. 2 for each circuit, and more extensive comparisons are shown in Supplementary Information Section S2.

We applied dynamic characterization both sequentially and simultaneously to the circuits introduced in Fig. 2, using both direct and ratiometric approaches. The inference graph for sequential direct characterization is shown in Fig. 3, while the graph for ratiometric characterization is shown in Fig. S1. Below we summarize the various stages of the inference graph in Fig. 3.

Autofluorescence. We started by characterizing cellular autofluorescence, to explicitly account for autofluorescence in circuits expressing eYFP and eCFP (Fig. 2A; Section S2.1). We pharmacologically perturbed the cells with EtOH, which above a threshold concentration led to very slow cell growth. We found that the per-cell rate of autofluorescence production at 480 nm, corresponding to eCFP measurements, was an order of magnitude higher than at 530 nm, corresponding to eYFP measurements. The autofluorescence parameters a_{530} and a_{480} inferred from these experiments were reused in subsequent circuits.

Constitutive promoter. Since both eCFP and eYFP were used in most of the synthetic gene circuits we constructed, we characterized their stability using a simple circuit (Fig. 2B; Section S2.2) expressing the eCFP and eYFP fluorescent proteins constitutively using the PR promoter, and measured their fluorescence for 36 hours. A model of the PRPR circuit incorporated autofluorescence terms, which shared information with the Auto circuit in Fig. 2A. The the time-series measurements of PRPR enabled identification of the eCFP and eYFP degradation parameters d_{CFP} and d_{YFP}.

Receivers. The first HSL devices we characterized were the Receivers, since these are also required to characterize the Relay and Degrader devices. We used the double receiver circuits developed in (27), which act as a dual reporter of 3-oxo-C6-HSL (C6) and 3-oxo-C12-HSL (C12; Fig. 2C; Section S2.3). This is achieved by the constitutive expression of LuxR and LasR and the inducible expression of fluorescent reporter proteins under the control of PLux promoter variants selected for orthogonal response to regulators composed of C6 bound to LuxR or C12 bound to LasR. In this way, eCFP reports the concentration of C6 and eYFP reports the concentration of C12, both across a broad range of concentrations. Because different promoters and ribosome binding sites are upstream of the coding sequences for eCFP and eYFP , as compared to the PRPR circuit, the parameters for their maximal production rate were assumed to take on new values. However, the degradation parameters were assumed to be the same as for PRPR, since the degradation rate is an intrinsic property of the protein. We used four different double receiver circuits, each incorporating a unique combination of promoter and ribosome binding sites in their LuxR and LasR expression cassettes. As explained in previous work, this enabled the effect of variable intracellular concentrations of LuxR and LasR to be modeled and therefore characterized quantitatively, which was critical for reducing chemical and genetic crosstalk (27).

Relays. We next characterized LuxI and LasI, the synthases of C6 and C12 respectively. HSL relay devices were designed that inducibly express the synthase of one signal in response to the other. HSL is produced intracellularly but quickly diffuses into the culture media, changing the signal concentration to which all cells are responding. This means that the promoter response to the synthesized signal is inherently time-varying and therefore cannot be accurately characterized using static methods. The first relay uses the C6-responsive pLux76 promoter to drive expression of LasI (abbreviated as P76-LasI), leading to intracellular production of C12 and subsequent induced expression of eYFP via the pLas81 promoter (Figs. 2D & S14B). The second uses the C12 responsive pLas81 promoter to drive expression of LuxI (abbreviated as P81-LuxI), leading to intracellular production of C6 and subsequent induced expression of eCFP (Fig. S14A). Again, reusing component characterization of the previously described circuits, we modeled time-series measurements of the response to different concentrations of C6 and C12 (Fig. 2D & S15), enabling quantification of the ability of LuxI and LasI to synthesize HSLs.

Inducer. We wished to use an orthogonal inducible system to drive expression of an HSL-degrading enzyme (see 'Degrader' below) but first needed an understanding of the parameters governing induction. To this end we characterized a simple circuit in which the properties of arabinose induction of the promoter PBAD could be quantified, by simply using PBAD to drive eYFP expression (Figs. 2E & S17). Incorporating the eYFP autofluorescence parameter from our Standard device, our model of the PBAD circuit closely described the exper-

bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

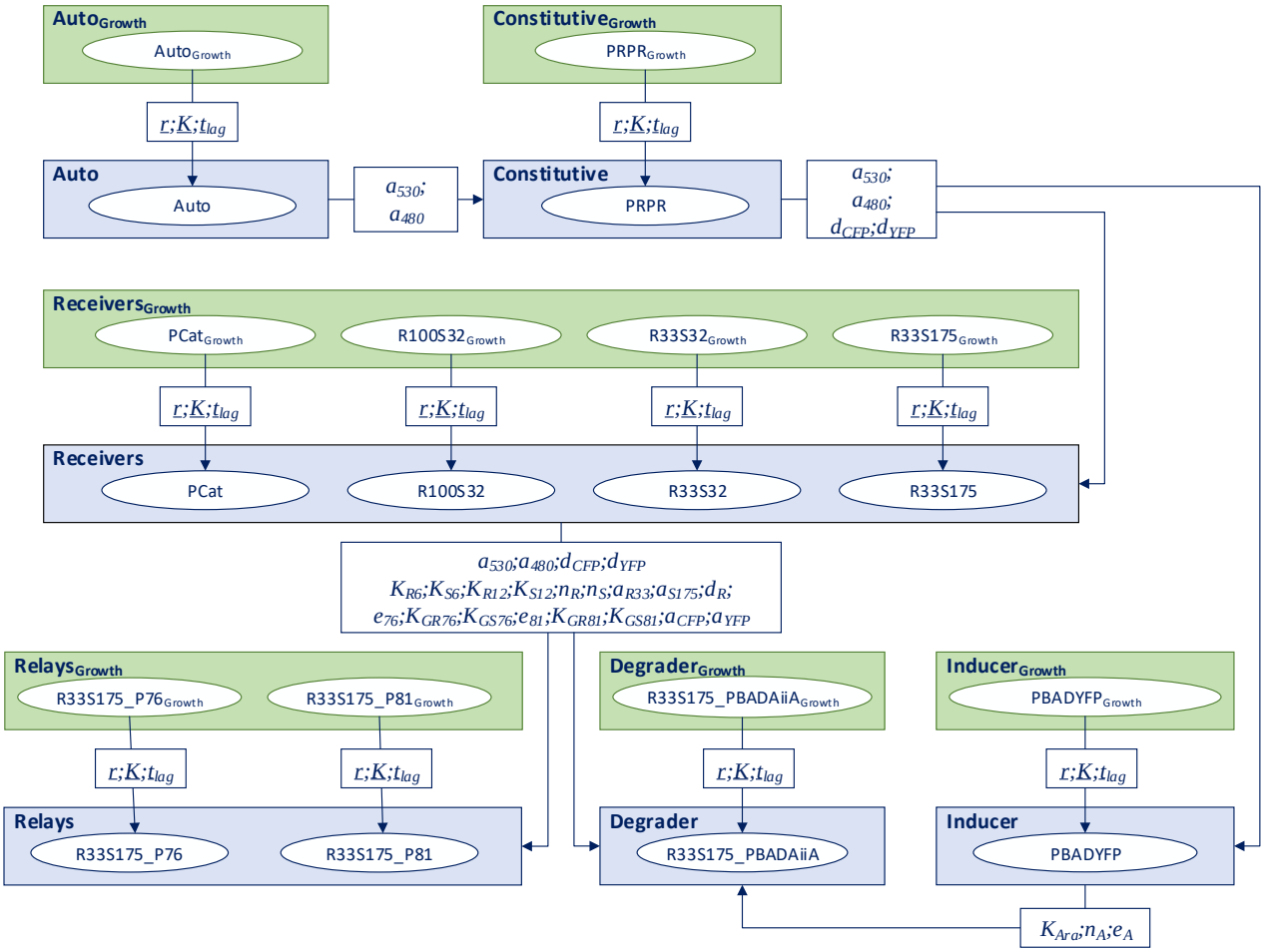

Fig. 3. Inference graph for direct dynamic characterization of synthetic gene circuits. For conciseness, we allow edges from the same source node to share common labels. The $\text{Auto}_{\text{Growth}}$ node uses a growth model to infer growth parameters r, K and t_{lag} of the autofluorescence circuit, under different treatments. The parameters are propagated as vectors of equal length (shown as underlined), where the length of the vectors corresponds to the number of treatments. The parameters are propagated to the Auto node, which uses the Auto model to infer parameters a_{530} and a_{480} that characterise circuit autofluorescence. The posterior distributions of these parameters are then used as priors for the Constitutive node, which uses a model of constitutive gene expression to infer the parameters d_{CFP} and d_{YFP}, together with updated posteriors of the autofluorescence parameters. The $\text{Receivers}_{Growth}$ node infers growth parameters of four receiver devices simultaneously. The devices differ primarily in the ribosome binding sites used to drive expression of the two receiver proteins, LuxR and LasR. The Receivers node uses the corresponding growth parameters to infer parameters shared between the four receiver circuits. The Relays, Degrader and Inducer nodes are defined in a similar fashion, with parameters propagated as shown in the inference graph.

imental measurements in response to different concentrations of arabinose (Fig. S18) and enabled identification of the transfer function parameters for arabinose-induction, given by K_{Ara}, n_{A} and e_{A} (Fig. S19).

Degrader. To characterize the HSL lactonase AiiA (28), we designed a circuit in which AiiA was placed under the control of the arabinose-inducible PBAD promoter (Fig. 2F), so that different intracellular levels of AiiA could be produced experimentally. Embedding the characterized PBAD and Receiver modules into the model of the AiiA circuit enabled modeling of the response to different concentrations of C6, C12 and arabinose. The parameters quantifying the shape of the transfer function of PBAD induction by arabinose were propagated from the characterization of the Arabinose circuit, though the parameter defining the maximal output ($a_{\text{YFP}}^{\text{PBAD}}$) was not, because that is associated with a different downstream protein coding region. Of particular importance in reproducing the observed bulk fluorescence dynamics of the Degrader circuit was the incorporation of the effect on cell growth of high arabinose induction of AiiA, which was observable in the OD_{600} data, and therefore quantified during the cell growth phase of dynamic characterization (Figs. S22 & S24).

Hypotheses about changes in host cell metabolism can be encoded in time-varying functions of gene expression capacity. As cells pass through different phases of growth, from lag to exponential to stationary, they undergo highly regulated changes in gene expression and metabolic state (29). These changes will have an effect on the expression of synthetic circuits and therefore on their dynamics. We observed such changes in gene expression in the eCFP and eYFP channels of

bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

our circuits that correlated with the onset of decreasing growth rate as measured by OD. Interestingly, such changes were much less prominent in the RFP channel (Fig. S28). We explored a number of hypotheses about the nature of those changes by evaluating the performance of different functional forms used to define gene expression capacity h_T.

Table 1. Functional forms used to define gene expression capacity $h_T(c)$. In all cases gene expression capacity $h_C(c,\xi_{ij},\phi)$ in the control phases uses the *Constant* hypothesis and its parameters $\xi_{ij} = r_{c_{ij}}$ are inferred during tthis phase. Additional parameters in $h_C(c,\xi_{ij},\theta)$ may differ, and are therefore inferred during the target phase.

Description	Target phase $h_T(c,\xi_{ij},\theta)$
Constant	$r_{c_{ij}}$
TargetGrowth Tracks the growth-rate γ, falling to a basal level ϵ_c.	$r_{c_{ij}}((1-\epsilon_c)\gamma(c)+\epsilon_c)$
TargetGrowthDelay Tracks the growth-rate γ at time $t-\tau$	$r_{c_{ij}}((1-\epsilon_c)\gamma(c(t-\tau))+\epsilon_c)$
TargetRSwitch Sigmoid that switches at $c = K_c$ from a well-specific value r_c to a relative value ϵ_c	$r_{c_{ij}}\frac{c^{n_c}+\epsilon_c K_c^{n_c}}{c^{n_c}+K_c^{n_c}}$
TargetSwitch Sigmoid that switches at $c = K_c$ from a well-specific value r_c to a non-well-specific basal value r_s	$\frac{r_{c_{ij}}c^{n_c}+r_s K_c^{n_c}}{c^{n_c}+K_c^{n_c}}$

We considered several contrasting yet relatively simple assumptions about the functional form of the gene expression capacity $h_T(c,\xi_{ij},\theta)$ for the target phase (see Table 1). The simplest assumption was that gene expression capacity remains constant through time. However, because circuit activity changes over time, this assumption may break down if host cell resources become limiting when circuit activity increases. Furthermore, as bacterial cultures transition from exponential to stationary growth phase, broad changes in gene expression may need to be accounted for. For this we considered four alternative time-dependent functions (Table 1). The best-performing function differed from circuit to circuit but overall, the time-varying functions performed better than a constant gene expression capacity 4.

In each case, the culture-specific component $r_{c_{ij}}$ was inferred during the control phase, under a *Constant* hypothesis. Then, the non-culture-specific parameters (e.g. ϵ_c in *TargetGrowth*) were inferred in the target phase alongside the target circuit parameters θ. As such, these hypotheses assume that the gene expression capacity for the chromosomally integrated control undergoes contrasting regulation from that of the plasmid-expressed devices. In principle, every gene expressed on the plasmid could potentially be regulated differently, which would require gene-specific functions or parameterizations. Here, we assume that the same function is applied to all genes in the target phase, to avoid a combinatorial increase in the number of parameters to infer.

Sequential inference enables scalable dynamic characterization. We compared the maximum log-likelihood score of the simultaneous characterization (Fig. 4H) with the sum of the maximum log-likelihood scores over the 6 sequential characterizations (Fig. 4G). This revealed a clear preference for sequential inference, most likely due to the flexibility afforded to model parameters when propagated from one inference problem to the next. To assess the extent to which parameters took advantage of this flexibility, we plotted the marginal distributions for each parameter resulting from inference in the 6 problems (Fig. S25 & S26). We found that the parameters associated with autofluorescence and fluorescent protein degradation were relatively stable as they were passed through the graph. Whereas parameters associated with the binding of LuxR/LasR to HSL signals and the modified PLux promoters varied by up to an order of magnitude or more. Our interpretation is that this reflects the inherent correlation between these parameters and the parameters associated with other HSL signaling circuits, characterized in inference problems downstream of the double receiver (see Fig. 3). The amount of variation also was greater in the TargetSwitch hypothesis than TargetGrowth hypothesis, which is likely due to TargetSwitch having more parameters associated with h_T to parameterize, and thus more potential for correlation and non-identifiability.

We found that for sequential inference, in general, ratiometric dynamic characterization performed better than direct dynamic characterization in the training phase (Fig. 4G), though this preference was reversed in the simultaneous strategy (Fig. 4H). When considering the breakdown of these scores across different nodes of the inference graph, we found that PRPR, the Relays and some hypotheses for the double receiver did not benefit from the ratiometric control, but in fact led to worse fits to the data. Ideally, the ratiometric approach should be at least as good as the direct approach, which is equivalent to the ratiometric control parameters all being to the same value. As such, the conclusions here suggest that the inferred ratiometric control parameters can inappropriately quantify gene expression capacity. Because this is the first attempt to incorporate a ratiometric control into dynamical models, and several simplifications have been made, we leave further investigation of how to more broadly specify ratiometric control parameters to a future research.

Dynamic characterization enables prediction of dynamic behavior. To test the ability of our modeling framework to predict dynamic behaviors, we measured the R33-S175 double receiver circuit in a regime previously not encountered in the training phase. The HSL treatments were applied at different times τ after the

bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

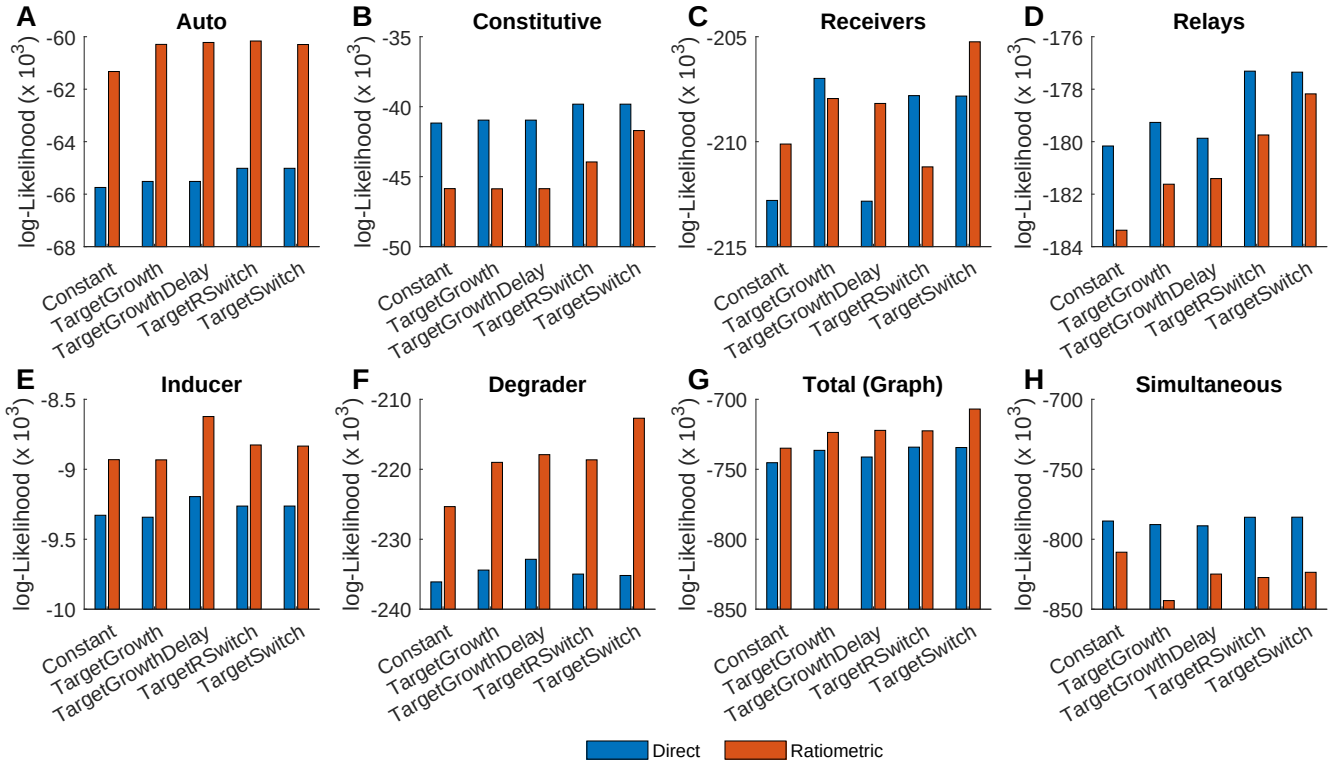

Fig. 4. Maximum-likelihood performance of dynamic characterization. Direct and ratiometric dynamic characterization was applied sequentially to synthetic gene circuits implementing (**A**) no expression (Auto), (**B**) constitutive expression (PRPR), (**C**) induced expression of HSL signals (DR), (**D**) relayed induced expression using HSLs (Relays), (**E**) arabinose-inducible expression (Inducer) and (**F**) disrupted HSL-inducible expression via arabinose-induced HSL degradation (AiiA). (**G**) Aggregated maximum likelihood score across A–F. (**H**) Maximum likelihood score of dynamic characterization applied simultaneously to the aforementioned circuits. In all cases, 5 different hypotheses about the dynamics of the gene expression capacity function were explored (see Table 1 for details).

initialization of the assay and the commencing of measurement. This allowed us to observe how otherwise untreated cells respond to step-like perturbations. As expected, there was a lag induced in the accumulation of eYFP and eCFP expression (Fig. 5).

To quantify the predictive capability of each model hypothesis, we ran step 1 (and step 2 for ratiometric) of the dynamic characterization method. We then propagated the cell density parameters (and gene expression capacity parameters as appropriate) to a target phase. However, in the target phase, rather than running MCMC-based parameter inference, we sampled the marginal parameter posterior distributions to produce a posterior predictive distribution (see Methods). We chose to marginalize at this point, as opposed to sampling from the joint posterior of the double receiver inference results. This is because in general, we would seek to predict circuits where the underlying parameters were characterized across a range of other circuits, which means that in this method, the statistical correlations would be unavailable for sampling. Marginalization will introduce an approximation error, most likely resulting in greater uncertainty, and so using it here provides a conservative estimate of the uncertainty in the predictive distribution. We also marginalized the parameter posteriors for predicting based on the simultaneous method, for fairness.

We performed two analyses of the posterior predictive distribution. First, we calculated the log-likelihood for each sample, and summarized these distributions for each model hypothesis considered (Fig. 5A). We then plotted the distribution of dynamical behaviors of the predictions for a selection of hypotheses. In Fig. 5B, we show the TargetSwitch hypothesis for the sequential application of direct dynamic characterization, which achieved a tight posterior predictive distribution. In this case, the general trend of most of the data is well reproduced. In Fig. S27, we show the TargetSwitch hypothesis again, but inferred from the simultaneous application of ratiometric dynamic characterization. Here, the predictions show a feature not exhibited by the measurements, which is that eYFP expression increases faster in the delayed HSL treatment case of $\tau = 4.82$ h, whereas the measurements show significantly reduced expression compared with $\tau = 0$ h and $\tau = 2.5$ h. We found that this was due to the ratiometric control (mRFP1 measurements) showing increased expression, which led to higher values of the r_c parameters, which when propagated to the double receiver model, result in higher rates of eYFP and eCFP production (Fig. S28).

Software implementation. Our method has been implemented in a dynamic characterization (DC) software package that enables a convenient specification of

bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

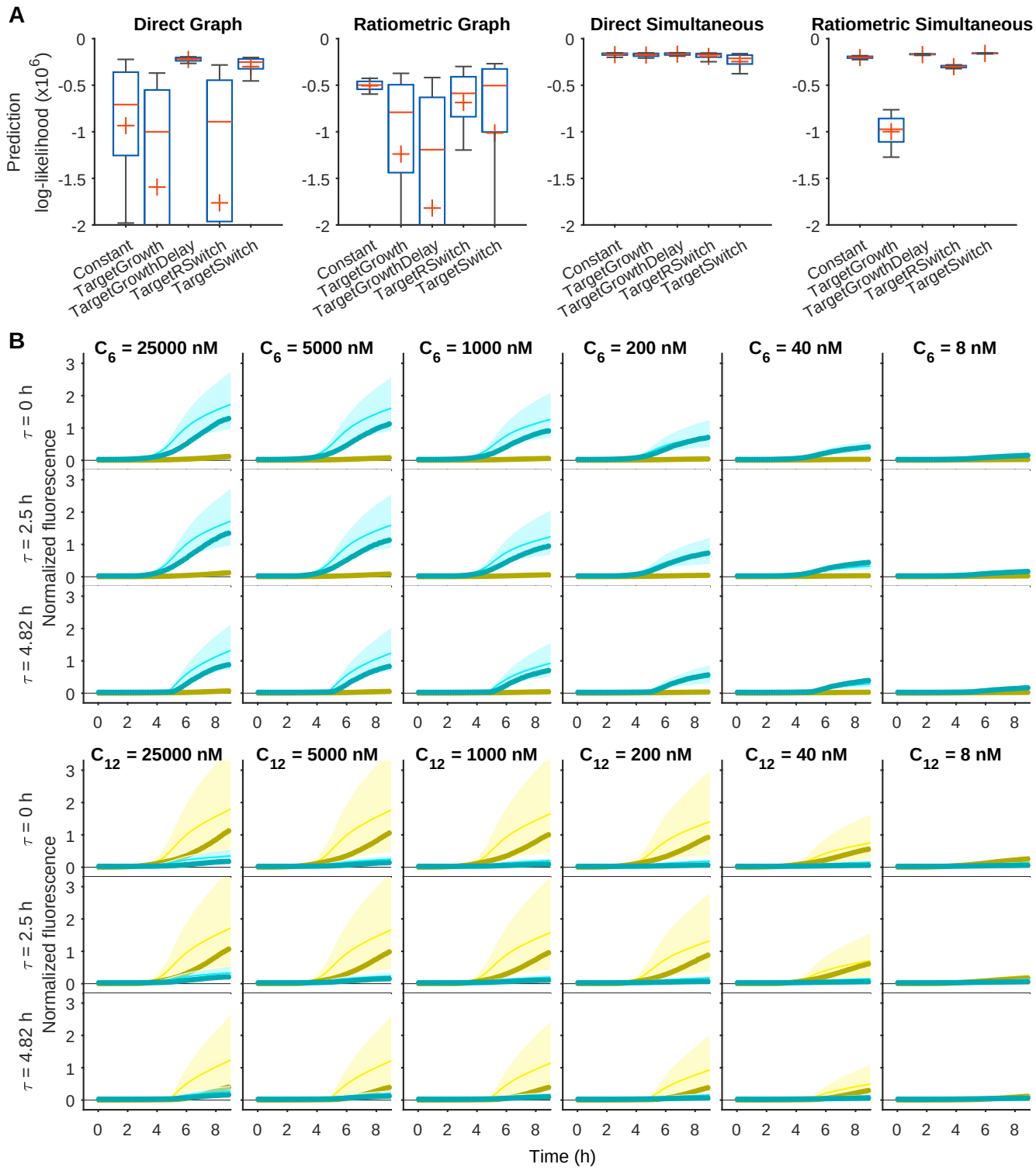

Fig. 5. Prediction of unseen data. The predictive capabilities of each model are compared against data measuring the R33-S175 double receiver device in response to HSL treatment at different time-points. **A.** The posterior predictive distribution is quantified for ratiometric and direct dynamic characterization, both using the graph-based and simultaneous methods, for the 5 hypotheses about the dynamics of gene expression capacity. The box-plots indicate the interquartile range of samples from the marginalized parameter posteriors. The red plus symbols indicate the means of each distribution. **B.** Comparison of the posterior predictive distribution for the TargetSwitch model using the direct graph method. In each column, the HSL treatment is indicated at the top, and in each row the time of HSL addition is noted on the left. eYFP and eCFP data are shown as thick darker lines (yellow and cyan respectively), with model simulations depicted as the mean (thin solid lines) and 95% credibility intervals as the shaded region.

model hypotheses using domain-specific programming languages genetic engineering of living cells (GEC) and chemical reaction networks (CRN). Software tools employing these languages already incorporate MCMC-based parameter inference for CRNs that are simulated under ODE semantics, and have been used in other

bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

research (30, 31). Here, we implemented an inference graph framework that supports both simultaneous and sequential dynamic characterization, and incorporated this into the GEC software tool. We also developed a wrapper program that can extract time-series data from a format that is produced from BMG plate-readers, and sub-select those wells that contain measurements for a device of interest. This enables a user to specify which devices they are interested in, provide links to the available data, and to generate and evaluate a set of inference problems that implements dynamic characterization methods. After each phase of the method, results files are written that enable the computation to be paused and resumed (check-pointing), so that a dependency graph can be navigated conveniently. Detailed information about how to use the dynamic characterization software is provided in Supplementary Information Section S1.

Discussion

Quantitative biological knowledge is difficult to systematize as it is not always clear how each measured quantity relates to others when factors such as genetic context and measurement techniques vary between experiments. Mathematical models provide a framework for explicitly encoding those relationships, allowing an internally consistent body of knowledge to be built, even when the systems being investigated are too large and complex to reason about directly. The use of mathematical models to make design decisions in the construction of synthetic gene circuits has been a central goal of the field of synthetic biology since its inception (6, 7). To make this a viable approach, methods are required that allow the parameters of models to take on values that cause models to match experimental data and thereby encode the quantitative biological knowledge gained by conducting experiments.

Our methods allow us to build increasingly complex genetic circuits by reusing genetic parts in new contexts and predicting their behavior by encoding our knowledge from previous experiments in mathematical models parameterized by the data from those experiments. By taking a Bayesian approach we also capture our uncertainty about the values those parameters can take and we can use that uncertainty measure to determine which experiments to perform next.

Dynamic characterization methods allow for quantitative understanding of inherently time-varying circuits such as HSL senders and degraders. Most methods that characterize reporter gene expression assume the reporter to be in a steady state due to production, dilution, and degradation. While this simplification can be justified in many synthetic circuits, it is distinctly unsuitable for circuits that produce or degrade intercellular signals while responding to those signals. In a culture of one of these circuits, the population of cells will be growing and continuously producing (or degrading) signal, meaning that the signal concentration, and the resultant reporter expression in response, will be changing throughout the experiment. Our approach was designed to capture these dynamic effects and therefore allows a direct quantitative characterization of these circuits for the first time.

In our experiments with the HSL degrading enzyme AiiA, inducing expression even at fairly moderate levels in combination with a receiver circuit results in strong effects on growth and gene expression capacity (Figs. S22 & S24). These effects mean attempting to determine the efficacy of degradation under various induction conditions directly from the raw data is extremely difficult. By applying our method, which can take into account both a decreased growth rate and altered gene expression capacity, we learn the rate at which HSLs are degraded for a particular level of expression of AiiA, which is consistent with both the PBAD induction measured with a YFP reporter and with our characterization of the receiver device.

Sequential versus simultaneous inference. Our inference graph methodology supports both sequential and simultaneous inference, where simultaneous inference involves running a single inference problem over all circuits simultaneously. There are benefits and drawbacks to each.

Sequential inference enables scalable parameter inference with respect to a growing collection of genetic circuits and associated measurements. This is particularly important for circuits with large numbers of parameters, including culture-specific parameters that account for variations in cell growth and gene expression capacity. Sequential inference is achieved by propagating marginal parameter posteriors of one inference problem as the priors of another, for example using truncated Gaussian distributions (Fig. 1B-D), or by propagating maximum likelihood parameter estimates directly. Performing inference sequentially in multiple phases allows each phase to be computationally tractable. Since scientific investigation proceeds in iterations of hypothesis formulation and testing, sequential inference provides a practical solution for updating model parameters as new data arrive. In contrast, simultaneous inference can only incorporate new data by re-running inference over all existing data for all existing circuits. This becomes intractable as the number of circuits increases, due to the increased number of models to simulate and the number of samples needed for convergence.

Sequential inference enables model simulations to more closely match experimental observations, since there is flexibility of parameters between circuits. However, this increased flexibility comes at the cost of increased uncertainty, which we observed to be greater in the posterior predictive distribution of unseen data

bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Fig. 5A). In contrast, simultaneous inference requires parameter values to be identical across all datasets, which can sometimes be difficult to reconcile.

Simultaneous inference enables parameters shared between multiple genetic circuits to be identified in a non-biased way. In contrast, sequential inference introduces two approximation errors, the first through marginalization of the posterior parameter distributions, and the second through approximation of these marginals with truncated Gaussian distributions, which may not be the optimal choice of parametric distribution in some cases. Therefore, one must be careful to only marginalize over MCMC chains that have reached the same local maximum of the likelihood function, and discard those stuck in other local maxima. In the case of propagating maximum likelihood estimates for parameters, such as cell growth and gene expression capacity, this leads to an under-representation of parameter uncertainties, which can in turn lead to inaccuracies in a Bayesian treatment of the target phase.

Open versus closed loop modeling. Our models of genetic circuits assume a particular level of abstraction for describing the time-evolution of constituent components, by describing transcription and translation explicitly but ignoring host cell machinery (RNA polymerase, ribosomes, proteases, etc). More detailed models could offer an alternative starting point, but without measurements corresponding to careful manipulation of the host machinery, such models are likely to contain non-identifiable parameters and misspecifications. Without explicit consideration of host cell resources, the way we have chosen to account for the feedback burden of circuit activity is open loop: each culture is modeled with independent parameters quantifying the rates of cell growth and gene expression. By breaking the loop in this way, our approach removes the need to speculate on the functional form of the feedback between a genetic circuit and the host cell. Such speculation would increase the complexity of models and increase the possibility of the feedback models compensating for misfits in the genetic circuit models, with poor mechanistic justification. Greater mechanistic understanding of the interaction between synthetic genetic circuits and cellular burden would likely improve both parameterizations and predictions, but in the absence of such insight, open-loop arrangements such as the one proposed here might be preferable to truly closed-loop feedback models.

There are drawbacks of this kind of simplification, however. Most notably, transient changes in burden that are associated with changes in circuit activity may not be captured. This is the situation encountered in experiments with timed addition of HSL. In these experiments we might expect burden to increase dramatically at the point in time that HSL is added to the culture. A predefined function for gene expression capacity will not objectively take into account this switch in feedback burden. Indeed, the direct dynamic characterization method, which has no inter-culture variability in gene expression capacity, produces a better fit to the timed addition experiment (Fig. 5). Curiously, the RFP signal in these experiments is larger in later treatments but still well fit by a constant gene expression capacity Fig. S28, which is not necessarily predicted by simple intuitions about the timing of burden. We also found that the direct dynamic characterization method scored more highly than the ratiometric method for PRPR, the Relays and sometimes the double receiver, suggesting that gene expression capacity is not well accounted for by the ratiometric control in certain cases. One possibility is that in general, chromosomal gene expression might differ from plasmid-derived gene expression. Alternatively, it might be that different circuits induce different burdens on their hosts, which may or may not be observable in a single ratiometric control signal.

Related and future work. Recent years have seen rapid improvement in Bayesian inference methodologies. Harnessing such improvements is a promising avenue for improving the scalability of genetic circuit characterization. For example, using stochastic gradient descent can enable more efficient approximate Bayesian inference (32). Beyond this, making explicit approximations about the distributions of parameter uncertainties (e.g. variational inference) can lead to efficient and more scalable inference schemes in the context of ODEs (33).

The simple gene expression capacity functions we considered do not fully capture the complexity of the metabolic shift that occurs in the transition of a cell culture from exponential to stationary phase. This is evidenced by their inability to exactly match the shape of the eCFP and eYFP time courses. Future work could improve this fit by including more mechanistic understanding about the way gene expression is affected by metabolic shifts, or by taking a more 'black box' approach to generating a functional form that best fits the data (33). Machine learning approaches that combine mechanistic models where such mechanism is known with learned functions that describe unknown mechanisms seem uniquely suited for this kind of problem.

Previous methods that characterize reporter gene expression typically assume the reporter to be in a steady state due to production, dilution, and degradation. While this simplification can be justified in many synthetic circuits, it is distinctly unsuitable for circuits that produce or degrade intercellular signals while responding to those signals. In a culture of one of these circuits, the population of cells will be growing and continuously producing (or degrading) signal, meaning that the signal concentration, and the expression of reporter responding to the signal, will be changing throughout the experiment. Our approach was

bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

designed to capture these dynamic effects, allowing a quantitative characterization of these circuits for the first time.

Our method allows us to build increasingly complex genetic circuits by reusing genetic devices and predicting their behavior, and to encode our knowledge from previous experiments in mathematical models parameterized by data from those experiments. By taking a Bayesian approach we also capture our uncertainty about those parameters values, and could in future use this uncertainty to determine the next experiments to perform.

Methods

Measurement of synthetic gene circuits in a microplate fluorometer. Plate fluorometer assays were conducted as previously described (27). Briefly, overnight cultures of cells containing constitutive chromosomal mRFP1 (27) and the plasmid construct of interest were diluted 1:100, grown to an OD of approximately 0.5, then diluted 1:1000 into M9 supplemented with 0.2% casamino acids and 0.4% glucose. 200μl of culture was aliquoted into each well and measurements were taken every 10 min for 1,000-2,000 min in a BMG FLUOstar Omega plate fluorometer. 3-oxohexanoyl-homoserine lactone (C6-HSL, Cayman Chemicals) and 3-oxododecanoyl-homoserine lactone (C12-HSL, Cayman Chemicals) were dissolved to a concentration of 200 mM in DMSO then C6-HSL was diluted in supplemented M9 medium to the concentrations described, while C12-HSL, due to its limited solubility in aqueous media, was first diluted 1:50 in ethanol then diluted in supplemented M9 medium to the concentrations described. A 1M arabinose (Sigma) stock solution was made in water, filter sterilized, and diluted in supplemented M9 medium to the concentrations described. HSL receiver and sender plasmids were previously described in (27) and all other plasmids (summarized in Table S1) were constructed using Gibson Assembly (34) from parts obtained from the MIT Registry of Standard Biological Parts (`http://parts.igem.org`).

Observer process. Because the microplate reader measurements monitor a whole culture growing over time, the observed fluorescence measurements are a sum of the single cell fluorescences in the culture. However, our models of synthetic gene circuits describe intracellular interactions, and therefore describe concentrations within each cell. Therefore, to compare the models with bulk fluorescence data, we use a nonlinear *observer* process. For each wavelength w, we describe bulk fluorescence variables as

$$\begin{aligned} B_{610} &= ([\text{RFP}]) \times c + B_{610}^{\text{back}} \\ B_{480} &= ([\text{CFP}] + [F_{480}]) \times c + B_{480}^{\text{back}} \\ B_{530} &= ([\text{YFP}] + [F_{530}]) \times c + B_{530}^{\text{back}} \end{aligned} \quad (4)$$

where $[F_w]$ is the per-cell autofluorescence at wavelength w, c is cell density (as above), and B_w^{back} is the constant background fluorescence at wavelength w. We use [RFP], [CFP] and [YFP] to denote the (modelled) intracellular concentrations of mRFP1 , eCFP and eYFP respectively. Following these definitions, it remains to describe how to model the cell density (c) and the intracellular concentrations of fluorescent proteins ([RFP], [CFP] and [YFP]).

Modeling cell growth. We characterized the growth of each culture in a microtiter plate separately, to explicitly capture acceleration and deceleration of growth in response to variations in synthetic gene circuit activity and culture density. Absorbance measurements at 600 nm enable quantification of the optical density over time (OD_{600}), providing a measure that is proportional to cell density (35).

To model cell density, we use the general equation

$$\dot{c} = \gamma(c;\rho).c \quad (5)$$

where $\gamma(c;\rho)$ is the *specific growth rate* function that is parameterized by ρ. For the lag-logistic model, γ is defined by

$$\gamma(c;\rho) = \begin{cases} 0, & t < t_{\text{lag}} \\ r\left(1 - \dfrac{c}{K}\right), & t \geq t_{\text{lag}} \end{cases} \quad (6)$$

Here r is the per capita culture growth rate, K is the carrying capacity representing the maximum culture density and t_{lag} is the duration of the lag phase of bacterial growth. Accordingly, the parameters of cell growth for circuit i and culture j are given by $\rho_{ij} = \{r_{ij}, K_{ij}, t_{\text{lag},ij}\}$. The time-evolution of $c(t)$ depends also on the initial cell density, $c^0 := c(t=0)$, which is fixed to the intended dilution factor (0.2%) during preparation of the assay. In the measured OD_{600} signal, there is a non-cellular background signal, which we parameterize as a plate-level shared parameter x_b. As a closed-form solution exists for the lag-logistic model, we can simply model the time evolution of the OD signal in well i as

$$\hat{x}_{ij}(t) = \frac{K_{ij} e^{r_{ij}(t - t_{\text{lag}})} c^0}{K_{ij} - c^0 + e^{r_{ij}(t - t_{\text{lag},ij})} c^0} + x_b \quad (7)$$

By assuming that the initial cell density is small, we have that $c_i(0) \approx 0$, and so the background absorbance measurement will be approximately equal to the initial absorbance measurement. Therefore, we identify the parameter x_b as the average over the first time point of the OD signal in each well.

This leaves the identification of the well-specific parameters as a set of inference problems in which we seek to minimize the deviation between the data (x_k) and simulation ($\hat{x}(t_k)$). To obtain parameter estimates, we assume that the deviation between data and simulation is Gaussian-distributed with standard deviation

bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

σ. We seek to maximize the log-likelihood of producing the data in well i with parameters ρ_i

$$\max_{\rho_i}\{L(\rho_i)\} \tag{8}$$

where $L(\rho_i)$ is the sum of the log-probabilities of Gaussian observations for $x_k - \hat{x}_k(t_i)$, with variance σ^2. For simple optimization problems over few parameters, it is possible to obtain good performance (both reliability and efficiency) using direct search methods such as the Nelder-Mead simplex algorithm (36). Here, we used the implementation in MathNet.Numerics (http://numerics.mathdotnet.com/) throughout this study for cell growth characterization.

Control phase model derivation. We describe constitutive production of mRFP1 mRNA, followed by translation and fluorescent protein maturation by the following CRN:

$$\begin{aligned} g_{\text{RFP}} &\xrightarrow{k_m} g_{\text{RFP}} + m_{\text{RFP}} \\ m_{\text{RFP}} &\xrightarrow{k_p} m_{\text{RFP}} + \text{iRFP} \\ \text{iRFP} &\xrightarrow{k_{\text{mat}}} \text{RFP} \\ m_{\text{RFP}} &\xrightarrow{d_m+\gamma(c)} \emptyset \\ \text{iRFP} &\xrightarrow{d_{\text{RFP}}+\gamma(c)} \emptyset \\ \text{RFP} &\xrightarrow{d_{\text{RFP}}+\gamma(c)} \emptyset \end{aligned} \tag{9}$$

where k_m, k_p and μ_R are transcription, translation and maturation of mRFP1 respectively, γ is the rate of dilution, and d_m, d_{RFP} are degradation rates of mRNA and (mRFP1) protein.

From the reaction set in (Eq. 9), we can write down corresponding ordinary differential equations for the time evolution of the concentrations of each molecule, assuming mass action kinetics as follows:

$$\begin{aligned} [\dot{m_{\text{RFP}}}] &= k_m - (d_m + \gamma(c))[m_{\text{RFP}}] \\ [\dot{\text{iRFP}}] &= k_p[m_{\text{RFP}}] - (d_{\text{RFP}} + \gamma(c) + \mu_R)[\text{iRFP}] \\ [\dot{\text{RFP}}] &= \mu_R[\text{iRFP}] - (d_{\text{RFP}} + \gamma(c))[\text{RFP}] \end{aligned} \tag{10}$$

The dynamics of transcription and mRNA turnover are usually faster than translation and protein turnover, and therefore a separation of timescales argument can be applied to simplify the model equations (Eq. 10) as $[m_{\text{RFP}}]^* \approx {}^{k_m}/_{d_m+\gamma}$. By further assuming that the mRNA dynamics are faster than dilution, the above expression becomes constant, and can be incorporated into a single quantity incorporating the mRNA equilibrium and the translation rate k_p.

$$[\dot{\text{iRFP}}] = k_c - (d_{\text{RFP}} + \gamma(c) + \mu_R)[\text{iRFP}] \tag{11a}$$

$$[\dot{\text{RFP}}] = \mu_R[\text{iRFP}] - (d_{\text{RFP}} + \gamma(c))[\text{RFP}] \tag{11b}$$

If the maturation time is faster than the turnover time for the fluorescent proteins, then an even simpler model can be derived:

$$[\dot{\text{RFP}}] = a_{\text{RFP}} - (d_{\text{RFP}} + \gamma(c))[\text{RFP}] \tag{12}$$

where a_{RFP} incorporates transcription, translation, fluorescent protein maturation and mRNA degradation rates. Since these factors are might be modulated by the growth phase and metabolic status of cell, a_{RFP} can be considered to represent the gene expression capacity, as introduced above. Accordingly, we introduce the open loop parameterization of the control circuit as

$$[\dot{\text{RFP}}] = h_C(c) - (d_{\text{RFP}} + \gamma(c))[\text{RFP}] \tag{13}$$

where gene expression capacity $h_C(c, \xi_{ij}, \phi)$ is parameterized by culture-specific parameters ξ_{ij} and global parameters ϕ (see Table 1 for options). The effect of dilution is described by the specific growth rate $\gamma(c)$ (Eq. 6) and represents the decline in concentration as the volume of cells increases. Since we have characterized cell growth for each culture, we can use this to accurately model culture-specific dilution.

Target phase models. Here, we summarize the derived equations for each (target phase) circuit considered in this study. Full derivation of these models can be found in Supplementary Information Section S2.

Auto . The model of autofluorescence assumes that each cell produces material autofluorescent at the wavelengths corresponding to YFP and CFP measurement, and is subject to growth dilution in the usual way.

$$[\dot{F_{530}}] = h_T(c).a_{530} - \gamma(c).[F_{530}] \tag{14a}$$

$$[\dot{F_{480}}] = h_T(c).a_{480} - \gamma(c).[F_{480}] \tag{14b}$$

Constitutive . The model of the Constitutive circuit incorporates autofluorescence (Eq. 14) and the following equations describing constitutive expression, degradation and growth dilution of eYFP and eCFP .

$$[\dot{\text{YFP}}] = h_T(c).a^{\text{PR}}_{\text{YFP}} - (d_{\text{YFP}} + \gamma(c)).[\text{YFP}] \tag{15a}$$

$$[\dot{\text{CFP}}] = h_T(c).a^{\text{PR}}_{\text{CFP}} - (d_{\text{CFP}} + \gamma(c)).[\text{CFP}] \tag{15b}$$

Receivers . The receivers are a collection of 4 different circuits which differ in their ribosome binding sites upstream of LuxR and LasR expression. Accordingly, their models all have the same structure, but use RBS-

bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

specific parameters for LuxR and LasR synthesis.

$$[\dot{\mathrm{R}}] = h_T(c).a_R - (d_R + \gamma(c)).[\mathrm{R}] \quad \textbf{(16a)}$$

$$[\dot{\mathrm{S}}] = h_T(c).a_S - (d_R + \gamma(c)).[\mathrm{S}] \quad \textbf{(16b)}$$

$$[\dot{\mathrm{CFP}}] = h_T(c).a^{\mathrm{P76}}_{\mathrm{CFP}}.f_{76}([\mathrm{C}_6],[\mathrm{C}_{12}],[R],[S]) - (d_{\mathrm{CFP}} + \gamma(c)).[\mathrm{CFP}] \quad \textbf{(16c)}$$

$$[\dot{\mathrm{YFP}}] = h_T(c).a^{\mathrm{P81}}_{\mathrm{YFP}}.f_{81}([\mathrm{C}_6],[\mathrm{C}_{12}],[R],[S]) - (d_{\mathrm{YFP}} + \gamma(c)).[\mathrm{YFP}] \quad \textbf{(16d)}$$

As above, the full models also incorporate the autofluorescence equations (Eq. 14).

Relays . Two relay circuits are considered that each extend the R33-S175 Receiver circuit. The relay device P76-LasI expresses the C12 synthase LasI in response to C6 (via the P76 promoter), while the P81-LasI device expresses the C6 synthase LuxI in response to C12 (via the P81 promoter). Accordingly, their models have a common structure

$$[\dot{L}] = h_T(c).f_k([\mathrm{C}_6],[\mathrm{C}_{12}],[\mathrm{R}],[\mathrm{S}]) - (d_L + \gamma(c)).[L] \quad \textbf{(17a)}$$

$$[\dot{\mathrm{C}_{12}}] = k'_{\mathrm{C12}}.c.[L] \quad \textbf{(17b)}$$

where k is P76 or P81, and L represents LuxI or LasI. Combining these equations with the equations for the R33-S175 circuit completes the description.

Inducer . Arabinose-inducible expression of YFP uses a Hill regulation function f_{PBAD}, with a maximal transcription rate that differs from maximal transcription of YFP of other promoters.

$$[\dot{\mathrm{YFP}}] = a^{\mathrm{PBAD}}_{\mathrm{YFP}}.f_{\mathrm{PBAD}}(\mathrm{Ara}) - (d_{\mathrm{YFP}} + \gamma(c)).[\mathrm{YFP}] \quad \textbf{(18)}$$

Combining these equations with the Auto circuit model completes the description.

Degrader. The model of Arabinose-inducible AiiA uses the same Hill function as for the Inducer circuit, but maximal production of AiiA is divided out, as its concentration is not observed.

$$[\dot{A}] = f_{\mathrm{PBAD}}(\mathrm{Ara}) - (d_A + \gamma(c)).[A] \quad \textbf{(19a)}$$

$$[\dot{\mathrm{C}_6}] = -d_{A6}.c.\frac{[A].[\mathrm{C}_6]}{1 + K_{A6}.[\mathrm{C}_6] + K_{A12}.[\mathrm{C}_{12}]} \quad \textbf{(19b)}$$

$$[\dot{\mathrm{C}_{12}}] = -d_{A12}.c.\frac{[A].[\mathrm{C}_{12}]}{1 + K_{A6}.[\mathrm{C}_6] + K_{A12}.[\mathrm{C}_{12}]} \quad \textbf{(19c)}$$

Combining these equations with the equations for the R33-S175 circuit completes the description.

Markov chain Monte Carlo. In both the control phase and the target phase, we seek to infer the parameters of ODE models, given some observational data. For notational convenience, we define θ to be the vector of parameter values sought. In contrast to the optimization procedure in the cell growth phase, we use Markov chain Monte Carlo (MCMC) methods because we found that the Nelder-Mead algorithm performed poorly for inference problems over many parameters, frequently getting stuck in sub-optimal global optima of the cost function. MCMC methods are able to reduce, but not completely remove, the impact of this problem. MCMC methods also have the advantage of characterizing the uncertainty of parameter estimates, which can arise from several sources: measurement error, process error (molecular stochasticity) and model misspecification.

We used the Metropolis-Hastings algorithm as implemented in the Filzbach software (http://www.github.com/predictionmachines/filzbach) to perform MCMC parameterizations. This requires specifying a function that evaluates the log-likelihood score for a candidate parameter set θ, and prior distributions of each parameter, which encode our prior belief of its plausible values.

If we denote by $y_{w,i,j}$ the bulk fluorescence measurements at wavelength w ($w \in \{480, 530, 610\}$) in well i ($i = 1, \ldots, n_c$) at time-points t_j ($j = 1, \ldots, n_t$), then we can define a log-likelihood function for wavelength w as

$$L_w(\theta) = \log p(y_w|\theta) = -\sum_{i=1}^{n_c}\sum_{j=1}^{n_t}\frac{(B_w^{(i)}(t_j) - y_{w,i,j})^2}{2\sigma^2} - n_c n_t \log(\sigma\sqrt{2\pi})$$

where n_c is the number of cultures (wells in the 96-well plate). In the control phase, we simply use L_{610} as the complete likelihood function, while in the target phase we use $L_{480} + L_{530}$ when both eCFP and eYFP reporters are present in a circuit, and just one term otherwise. The parameter σ describes the standard deviation of the data. Here, we infer σ during application of MCMC.

Prior distributions encode our prior belief about the values of parameters. When characterizing a variable that has not previously been used in a model before, it can be difficult to know how to set the prior, so in this case we use a uniform distribution with wide bounds. This prevents the MCMC sampler from being swayed by an inappropriate value. When a parameter has been seen in another analysis, we propagate the marginal posterior from the previous analysis as a prior. To do this, we use truncated Gaussian distributions, where the mean and standard deviation are calculated from the previous MCMC samples, and the bounds are taken as the uniform prior bounds of that same analysis.

Computing the posterior predictive distribution. To evaluate a model against unseen data, we approximate the posterior predictive distribution of the data, given our best estimates of the distributions of the models parameters. To do this, we formulate a model of the synthetic gene circuit being measured. In this article,

bioRxiv preprint doi: https://doi.org/10.1101/635672; this version posted May 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

we show calculations for the posterior predictive distribution of a circuit for which the parameters have already been characterized, so all are specified as truncated Gaussians, as described above. The predictions are formed by applying the cell growth phase and control phase as appropriate for dynamic characterization, but for the target phase we integrate over the prior. Accordingly, we approximate $p(y'|y)$ by marginalizing over the posteriors of $p(\theta|y)$ to give approximate priors π_θ and then producing Monte Carlo samples $\theta_k \sim \pi_\theta$, as

$$\log p(y'|y) = \log \int p(y'|\theta)p(\theta|y)d\theta$$
$$\approx \frac{1}{N}\sum_{k=1}^{N} \log p(y'|\theta_k)$$

ACKNOWLEDGEMENTS
The authors thank Fernan Federici and Timothy Rudge (Universidad Catolica de Chile) for useful discussions about ratiometric characterization. We also thank Ted Meeds (Microsoft Research) and Geoffrey Roeder (Princeton University) for discussions about Bayesian inference.

Bibliography

1. Ferdinand Sedlmayer, Dominique Aubel, and Martin Fussenegger. Synthetic gene circuits for the detection, elimination and prevention of disease. *Nature Biomedical Engineering*, 2 (6):399–415, 2018. ISSN 2157846X. doi: 10.1038/s41551-018-0215-0.
2. J. H. Collins and E. M. Young. Genetic engineering of host organisms for pharmaceutical synthesis. *Curr. Opin. Biotechnol.*, 53:191–200, 10 2018.
3. J. Huang, S. Liu, C. Zhang, X. Wang, J. Pu, F. Ba, S. Xue, H. Ye, T. Zhao, K. Li, Y. Wang, J. Zhang, L. Wang, C. Fan, T. K. Lu, and C. Zhong. Programmable and printable Bacillus subtilis biofilms as engineered living materials. *Nat. Chem. Biol.*, 15(1):34–41, 01 2019.
4. F. Moser, A. Espah Borujeni, A. N. Ghodasara, E. Cameron, Y. Park, and C. A. Voigt. Dynamic control of endogenous metabolism with combinatorial logic circuits. *Mol. Syst. Biol.*, 14(11):e8605, 11 2018.
5. Keisuke Watanabe, Shunichiro Kuramitsu, Avery D Posey Jr, and Carl H June. Expanding the therapeutic window for car t cell therapy in solid tumors: the knowns and unknowns of car t cell biology. *Frontiers in Immunology*, 9, 2018.
6. Jennifer A.N. Brophy and Christopher A. Voigt. Principles of genetic circuit design, 2014. ISSN 15487105.
7. George M. Church, Michael B. Elowitz, Christina D. Smolke, Christopher A. Voigt, and Ron Weiss. Realizing the potential of synthetic biology, 2014. ISSN 14710080.
8. J. R. Kelly, A. J. Rubin, J. H. Davis, C. M. Ajo-Franklin, J. Cumbers, M. J. Czar, K. de Mora, A. L. Glieberman, D. D. Monie, and D. Endy. Measuring the activity of BioBrick promoters using an in vivo reference standard. *J Biol Eng*, 3:4, Mar 2009.
9. L. Keren, O. Zackay, M. Lotan-Pompan, U. Barenholz, E. Dekel, V. Sasson, G. Aidelberg, A. Bren, D. Zeevi, A. Weinberger, U. Alon, R. Milo, and E. Segal. Promoters maintain their relative activity levels under different growth conditions. *Mol. Syst. Biol.*, 9:701, Oct 2013.
10. Timothy J. Rudge, James R. Brown, Fernan Federici, Neil Dalchau, Andrew Phillips, James W. Ajioka, and Jim Haseloff. Characterization of Intrinsic Properties of Promoters. *ACS Synthetic Biology*, 5(1):89–98, 2016. ISSN 21615063. doi: 10.1021/acssynbio.5b00116.
11. A. A. K. Nielsen, Bryan S Der, J Shin, P Vaidyanathan, V. Paralanov, E. A. Strychalski, D. Ross, Douglas Densmore, and Christopher A. Voigt. Genetic circuit design automation. *Science*, 352(6281):aac7341–aac7341, 4 2016. ISSN 0036-8075. doi: 10.1126/science.aac7341.
12. Thomas E Gorochowski, Irem Avcilar-Kucukgoze, Roel AL Bovenberg, Johannes A Roubos, and Zoya Ignatova. A minimal model of ribosome allocation dynamics captures trade-offs in expression between endogenous and synthetic genes. *ACS synthetic biology*, 5(7):710–720, 2016.
13. Yili Qian, Hsin Ho Huang, José I. Jiménez, and Domitilla Del Vecchio. Resource Competition Shapes the Response of Genetic Circuits. *ACS Synthetic Biology*, 6(7):1263–1272, 2017. ISSN 21615063. doi: 10.1021/acssynbio.6b00361.
14. Lorenzo Pasotti, Massimo Bellato, Michela Casanova, Susanna Zucca, Maria Gabriella Cusella De Angelis, and Paolo Magni. Re-using biological devices: A model-aided analysis of interconnected transcriptional cascades designed from the bottom-up. *Journal of Biological Engineering*, 11(1), 2017. ISSN 17541611. doi: 10.1186/s13036-017-0090-3.
15. Axel Nyström, Antonis Papachristodoulou, and Andrew Angel. A Dynamic Model of Resource Allocation in Response to the Presence of a Synthetic Construct. *ACS Synthetic Biology*, 7(5):1201–1210, 5 2018. ISSN 2161-5063. doi: 10.1021/acssynbio.8b00015.
16. Francesca Ceroni, Rhys Algar, Guy Bart Stan, and Tom Ellis. Quantifying cellular capacity identifies gene expression designs with reduced burden. *Nature Methods*, 12(5):415–418, 2015. ISSN 15487105. doi: 10.1038/nmeth.3339.
17. Evangelos-Marios Nikolados, Andrea Y Weisse, Francesca Ceroni, and Diego A Oyarzun. Growth defects and loss-of-function in synthetic gene circuits. *bioRxiv*, page 623421, 2019.
18. Thomas E. Gorochowski, Eric Van Den Berg, Richard Kerkman, Johannes A. Roubos, and Roel A.L. Bovenberg. Using synthetic biological parts and microbioreactors to explore the protein expression characteristics of escherichia coli. *ACS Synthetic Biology*, 3(3):129–139, 2014. ISSN 21615063. doi: 10.1021/sb4001245.
19. Cheemeng Tan, Philippe Marguet, and Lingchong You. Emergent bistability by a growth-modulating positive feedback circuit. *Nature Chemical Biology*, 5(11):842–848, 2009. ISSN 15524469. doi: 10.1038/nchembio.218.
20. Philippe Bouvet and Joel Belasco. Control of RNase E-mediated RNA degradation by 5'-terminal base pairing in E. coli. *Nature*, 360(6403):488–91, 1992. ISSN 0028-0836. doi: 10.1038/360488a0.
21. Matthew Scott, Carl W Gunderson, Eduard M Mateescu, Zhongge Zhang, and Terence Hwa. Interdependence of Cell Growth and Gene Expression: Origins and Consequences. *Science*, 330(6007):1099–1102, 11 2010. ISSN 0036-8075. doi: 10.1126/science.1192588.
22. Boyan Yordanov, Neil Dalchau, Paul K. Grant, Michael Pedersen, Stephen Emmott, Jim Haseloff, and Andrew Phillips. A computational method for automated characterization of genetic components. *ACS Synthetic Biology*, 3(8):578–588, 2014. ISSN 21615063. doi: 10.1021/sb400152n.
23. Ann M Stevens and EP Greenberg. Quorum sensing in vibrio fischeri: essential elements for activation of the luminescence genes. *Journal of Bacteriology*, 179(2):557–562, 1997.
24. M Schuster, ML Urbanowski, and EP Greenberg. Promoter specificity in pseudomonas aeruginosa quorum sensing revealed by dna binding of purified lasr. *Proceedings of the National Academy of Sciences*, 101(45):15833–15839, 2004.
25. Tal Danino, Octavio Mondragón-Palomino, Lev Tsimring, and Jeff Hasty. A synchronized quorum of genetic clocks. *Nature*, 463(7279):326–330, jan 2010. ISSN 0028-0836. doi: 10.1038/nature08753.
26. Frederick K Balagaddé, Hao Song, Jun Ozaki, Cynthia H Collins, Matthew Barnet, Frances H Arnold, Stephen R Quake, and Lingchong You. A synthetic escherichia coli predator–prey ecosystem. *Molecular systems biology*, 4(1):187, 2008.
27. Paul K Grant, Neil Dalchau, James R Brown, Fernan Federici, Timothy J Rudge, Boyan Yordanov, Om Patange, Andrew Phillips, and Jim Haseloff. Orthogonal intercellular signaling for programmed spatial behavior. *Molecular systems biology*, 12(1):849, 2016. ISSN 1744-4292. doi: 10.15252/msb.20156590.
28. Dali Liu, Jessica Momb, Pei W. Thomas, Aaron Moulin, Gregory A. Petsko, Walter Fast, and Dagmar Ringe. Mechanism of the quorum-quenching lactonase (AiiA) from Bacillus thuringiensis. 1. Product-bound structures. *Biochemistry*, 47(29):7706–7714, 2008. ISSN 00062960. doi: 10.1021/bi800368y.
29. J. M. Navarro Llorens, A. Tormo, and E. Martinez-Garcia. Stationary phase in gram-negative bacteria. *FEMS Microbiol. Rev.*, 34(4):476–495, Jul 2010.
30. Gourab Chatterjee, Neil Dalchau, Richard A Muscat, Andrew Phillips, and Georg Seelig. A spatially localized architecture for fast and modular dna computing. *Nature nanotechnology*, 12(9):920, 2017.
31. Denise SM Boulanger, Ruth Charlotte Eccleston, Andrew Phillips, Peter V Coveney, Tim Elliott, and Neil Dalchau. A mechanistic model for predicting cell surface presentation of competing peptides by mhc class i molecules. *Frontiers in immunology*, 9:1538, 2018.
32. Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.
33. Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, 2018.
34. D. G. Gibson, L. Young, R. Y. Chuang, J. C. Venter, C. A. Hutchison, and H. O. Smith. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, 6 (5):343–345, May 2009.
35. A. L. Koch. Turbidity measurements of bacterial cultures in some available commercial instruments. *Anal. Biochem.*, 38(1):252–259, Nov 1970.
36. J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 1965. ISSN 0010-4620. doi: 10.1093/comjnl/7.4.308.