Modeling Genetic Circuit Behavior in Transiently Transfected Mammalian Cells

Junmin Wang,[*,†] Samuel A. Isaacson,[‡] and Calin Belta[†]

The Bioinformatics Graduate Program, Boston University, Boston, MA, USA, and Department of Mathematics, Boston University, Boston, MA, USA

E-mail: dawang@bu.edu

Abstract

Binning cells by plasmid copy number is a common practice for analyzing transient transfection data. In many kinetic models of transfected cells, protein production rates are assumed proportional to plasmid copy number. The validity of this assumption in transiently transfected mammalian cells is not clear; models based on this assumption appear unable to reproduce experimental flow cytometry data robustly. We hypothesize that protein saturation at high plasmid copy number is a reason previous models break down and validate our hypothesis by comparing experimental data and a stochastic chemical kinetics model. The model demonstrates that there are multiple distinct physical mechanisms that can cause saturation. Based on these observations, we develop a novel minimal bin-dependent ODE model that assumes different parameters for protein production in cells with low versus high numbers of plasmids. Compared to a traditional Hill-function-based model, the bin-dependent model requires only one additional parameter, but fits flow cytometry input-output data for individual modules up to

[*]To whom correspondence should be addressed
[†]The Bioinformatics Graduate Program
[‡]Department of Mathematics

twice as accurately. By composing together models of individually-fit modules, we use the bin-dependent model to predict the behavior of six cascades and three feed-forward circuits. The bin-dependent models are shown to provide more accurate predictions on average than corresponding (composed) Hill-function-based models and predictions of comparable accuracy to EQuIP, while still providing a minimal ODE-based model that should be easy to integrate as a subcomponent within larger differential equation circuit models. Our analysis also demonstrates that accounting for batch effects is important in developing accurate composed models.

Keywords

synthetic biology, modeling, transient transfection

In synthetic biology, there has been an increased use of transfection systems in mammalian cells in recent years. One reason for this increase is that transfection enables the production of important biomedical-related proteins, which can only become biologically active within mammalian cells.[1–4] Transient transfection is a common method for the delivery of foreign genetic materials into mammalian cells.[5–7] The transfected genetic materials utilize the cells' innate transcriptional and translational machineries to get expressed. Transiently transfected genes are only expressed temporarily, and do not become integrated into the host's genome. Compared with stable transfection, transient transfection offers faster expression of transfected genes, with higher expression levels. It also has lower cytotoxicity and induces no mutagenesis.[3,8,9] It has been shown to be an effective technique for speeding up the screening of novel synthetic designs.[10] These properties have motivated the investigation of transient transfection in mammalian synthetic biology.[3,11]

Modern synthetic biology is inseparable from the computational models that guide the construction of synthetic networks.[12] One challenge in building such models for mammalian cells arises from the need for a more comprehensive understanding of the cellular mechanisms underlying the transfection system.[12,13] Another challenge is predicting the behavior of

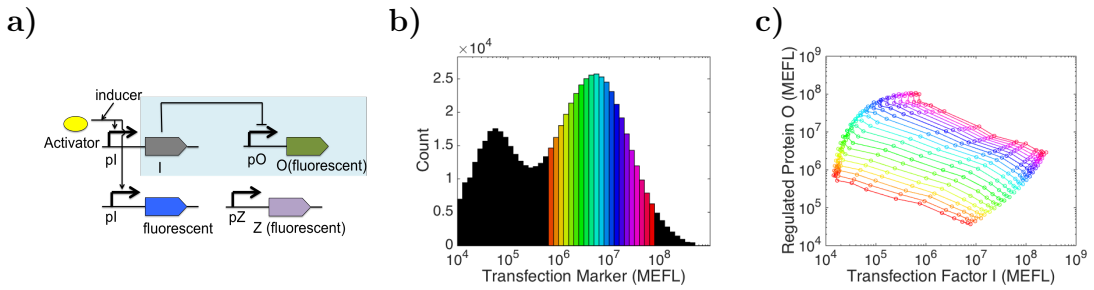


Figure 1: (a) Abstraction of a system comprised of a transfection marker and a module (in blue) encoding a transcriptional regulatory switch. See Supplementary Figure 1(a) and Ref. *26* for more detailed illustrations. The induced (input) gene I, activated by an inducer, regulates the expression of O, the regulated (output) gene. Z, the transfection marker, is used to estimate plasmid copy number. (b) Distribution of the transfection marker. The black bins are ignored because they represent untransfected cells (data from Ref. *26*). (c) Dose-response curves obtained from an experiment (data from Ref. *26*). Averaged measurements binned by the expression level of Z are shown by color. Cells are separated into bins of width 0.1 on a log scale. Each curve corresponds to a different bin. The 1st bin, represented by the curve at the bottom, contains cells with the lowest plasmid counts. Each dot represents the average concentrations of the induced protein and the regulated protein within a bin at a certain inducer level. Concentrations of the induced and the regulated proteins have units of MEFL. Details about data generation and binning can be found in Supporting Information Section 1.1.

genetic circuits based on the behavior of the building blocks of the circuits, also known as modules.[14–20] Chemical kinetic models have proven capable of describing circuit behavior in prokaryotic cells, which replicate foreign plasmids,[21,22] and in stably transfected eukaryotic cells in which plasmids are genome-integrated.[23] Plasmid copy number is assumed fixed in both of these scenarios. For transiently transfected mammalian cells (TTMC), there is a large variation in plasmid copy numbers across a population.[24,25] Binning cells by plasmid copy number is a common practice for analyzing flow cytometry data in this context (Figure 1(b)).[26–28] Subpopulations of cells with similar plasmid counts can then be studied in groups (Figure 1(c)). Developing a modeling approach that is compatible with binning is a prerequisite to building predictive models for complex circuits in TTMC. Davidsohn et al. developed a traditional Hill-function-based model for TTMC,[26] where the rate of protein production is assumed proportional to the average plasmid copy number in each bin. Unfortunately, as they demonstrated, this model does not fit their flow cytometry data well.

In this work, we hypothesize that high plasmid copy number may cause saturation in the levels of expressed proteins, leading to the breakdown of traditional Hill-function-based models in this context. To validate our hypothesis, we study detailed two-stage gene expression models of a transient co-transfection system via the Gillespie algorithm,[29,30] bin the simulated data by plasmid counts, and calculate the average protein concentrations within each bin. The agreement between the simulated results and the experimental data suggests that when physical gene expression parameters lie within a particular range, saturation of the rate of either transcription or translation can give rise to the observed saturated protein concentrations in experiments. These results suggest that the precise mechanism leading to the saturation of protein levels cannot be distinguished from just single-time flow cytometry measurements. To facilitate predictive modeling of circuits, we next develop a bin-dependent ordinary differential equation (ODE) model that splits flow cytometry data into two subsets based on plasmid copy number. This coarse-grained model can more accurately account for saturation in protein levels compared to standard Hill-function models, but avoids the need to specify a precise biological mechanism giving rise to saturation. For each plasmid copy number subset we fit separate kinetic parameters to the model, motivated by observations from the detailed stochastic model simulations. The resulting bin-dependent model is shown to outperform a traditional Hill-function-based model in reproducing input-output relationships for individual modules, yet requires only one additional parameter. By composing models fit to these *individual modules*, the bin-dependent model is also shown to predict the behavior of circuits composed of multiple modules more accurately than Hill-function-based models, while offering comparable accuracy to the EQuIP method of Ref. *26*. As the bin-dependent model is itself described by standard chemical-kinetics type ODEs for chemical concentrations, it can be easily integrated as a subcomponent within other differential equation circuit models, and easily extended to include more biological details or features for any given system. Note, in the remainder, species are denoted by Roman text, and concentrations by italicized text.

Results and Discussion

Experimental Data

The first step in building our circuit model is to examine experimental data. In this paper we adopt a bottom-up approach to making circuits via the assembly of individual modules, where a module is defined as a single transcriptional regulatory switch, consisting of a transcription factor, the downstream regulated promoter and its gene. As an example of the types of modules we will use, consider a module comprising a fluorescent-reporter system involving three fluorescent genes: the induced (input) gene, the regulated (output) gene and the transfection marker (Figure 1(a)). The expression levels of the fluorescent genes are measured via flow cytometry, with the fluorescence intensities used as proxies for the concentrations of the fluorescent proteins. The induced gene is regulated by a constitutive activator protein, and an external inducer whose concentration can be controlled. The product of the induced gene serves as a transcription factor for the regulated gene, controlling the latter's expression of a fluorescent reporter. The induced gene's product is not fluorescent, but is measured by co-expressing a fluorescent reporter gene of a different color from a promoter that has the same sequence but is encoded on a different plasmid.[31] The expression of the induced gene can be modulated by changing the amount of the inducer. Expression of the induced gene and the regulated gene at various inducer levels forms a dose-response curve (Figure 1(c)). In TTMC, expression levels are largely determined by the numbers of plasmids transfected in individual cells,[25,26] which cannot be controlled and are highly variable across a population. It is, therefore, necessary to estimate the plasmid copy numbers so that the effect of variation in copy numbers on gene expression can be captured. This is often achieved by co-transfecting another constitutively expressed fluorescent protein, which serves as the transfection marker (Figure 1(a)). The induced gene, the regulated gene, and the transfection marker can be encoded on either one plasmid or separate plasmids. The former ensures that there is a one-to-one correspondence among the genes. In comparison, the latter is often preferred as

separate plasmids can be absorbed by cells more readily due to smaller sizes, interference among the transcriptional units is minimized, and the concentrations of individual proteins can be adjusted more easily.[32,33] In what follows, we assume the transfection marker has been encoded on a separate plasmid for all models and experiments. We also assume the induced gene serves as an inhibitor of the regulated gene.

Fluorescence readings from flow cytometers can be converted to standard units of Molecules of Equivalent Fluorescein (MEFL) via TASBE Control.[26,34,35] Standardized data are segmented into bins by plasmid counts so that subpopulations of cells with similar plasmid counts can be studied in groups (Figure 1(c)).[26,27,36] Since flow cytometry measurements are typically log-normal distributed or a mixture of two log-normal distributions,[37,38] binning is performed on a log scale to ensure that each bin contains relatively equal numbers of cells. The width of bins is selected depending on the resolution at which analysis is to be conducted. An example of binning can be found in Supporting Information Section 1.1. In this paper, we will focus on the average temporal behavior within each bin, with the goal of developing ODE models that can be directly parameterized from binned flow cytometry data.

Protein Concentration vs Plasmid Copy Number

Hill functions are commonly used to model transcriptional regulation in ODE models (Figure 1(a)). (See Supporting Information Section 3 for a mathematical definition of a Hill function.) Davidsohn et al. developed a traditional Hill-function-based model to describe the time evolution of the induced and the regulated proteins in TTMC (Figure 1(a))[26,39] (see Supporting Information Section 3). A key assumption of their model is that the log of the maximal production rate of the regulated protein is a linear function of the log of the transfection marker. This assumption is supported by findings of several other studies in different biological contexts.[25,40] However, this assumption is only partially supported by the experimental data in Ref. *26*, shown here in Figure 2. When the induced gene is

minimally induced (0 nM of inducer), i.e., the regulated protein expressed without repressor, the log of the regulated protein's concentration grows proportionally to the log of the transfection marker between $10^{5.8}$ and 10^{7} MEFL for TAL14 and TAL21 or between $10^{5.8}$ and $10^{7.3}$ MEFL for LmrA. When the induced gene is fully induced (2000 nM of inducer), the log of the induced protein's concentration also grows linearly in the log of the transfection marker between $10^{5.8}$ and 10^{7} MEFL for TAL14 and TAL21 or between $10^{5.8}$ and $10^{7.3}$ MEFL for LmrA. Figure 2 also suggests that when either the induced gene or the regulated gene is maximally expressed, the concentrations of both the induced and the regulated proteins saturate starting from $10^{7.1}$ MEFL for TAL14 and TAL21 or $10^{7.4}$ MEFL for LmrA.

Furthermore, Figure 2 and the data in Ref. *26* suggest that when the induced gene is induced at 0nM, the log of the induced protein's concentration is near-constant for low plasmid copy numbers.[26] When the induced gene is fully induced, i.e., the regulated protein fully repressed, the log of the regulated protein's concentration grows linearly across all bins.

We now develop a detailed stochastic model of the plasmid system, similar to the one Davidsohn et al. constructed experimentally.[26] This model will enable us to explore possible mechanisms contributing to the observed saturation of protein concentrations at high plasmid copy number, as well as the near constant protein concentrations at low plasmid copy number. We do not attempt to fit this model to the single-time flow cytometry data directly as it is too complex to fit accurately without the incorporation of additional experimental measurements. Instead, *our purpose here is to use the stochastic model to gain a qualitative understanding of which biological hypotheses, and what ranges of physical gene expression parameters, may contribute to the observed saturation effect.* Our ultimate goal is to develop a simple model that qualitatively describes our limited set of data, avoiding further time-intensive experimental assays. Therefore, in the next subsection, we develop a more simplified ODE model that can be parameterized from just the limited flow cytometry data, building from the qualitative understanding of the two-plasmid system our stochastic model provides.

In our stochastic model, cells are co-transfected by a mixture of induced gene plasmids

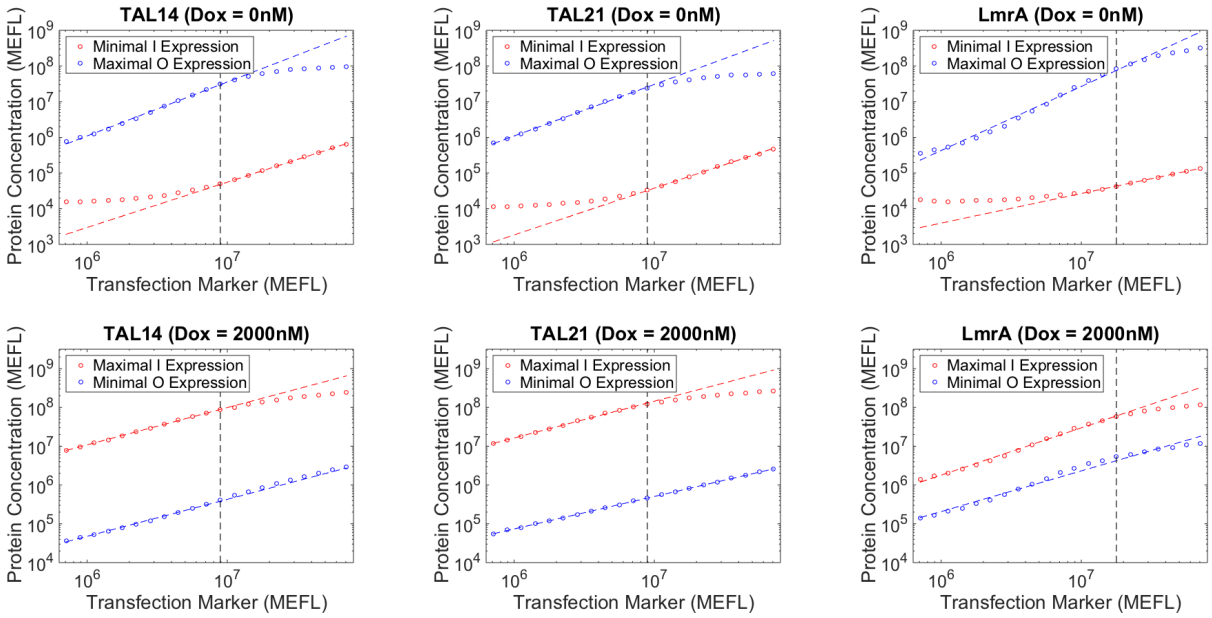


Figure 2: Maximal and minimal expressions of the induced gene I and the regulated gene O for TAL14, TAL21, and LmrA. In the figures, the x-axis corresponds to the concentration of the transfection marker, and the y-axis to the concentration of the input and the output proteins (here concentrations are in units of MEFL). Shown in red is the induced gene I, and in blue the regulated gene O. Each dot is the average protein concentration of cells from one bin. On the top row the circuit is induced at 0nM; on the bottom row, 2000nM. On the top row, least squares regression lines are fit to red dots from 10^7 to $10^{7.9}$ MEFL (TAL14 and TAL21) or from $10^{7.3}$ to $10^{7.9}$ MEFL (LmrA), and to blue dots from $10^{5.8}$ to 10^7 MEFL (TAL14 and TAL21) or from $10^{5.8}$ to $10^{7.3}$ MEFL (LmrA). On the bottom row, least squares regression lines are fit to red dots from $10^{5.8}$ to 10^7 MEFL (TAL14 and TAL21) or from $10^{5.8}$ to $10^{7.3}$ MEFL (LmrA), and to blue dots from $10^{5.8}$ to $10^{7.9}$ MEFL. The dots are calculated from the flow cytometry data of Ref. *26*.

and transfection marker plasmids. We focus on the dynamics of the transfection marker and the induced gene, which are integrated on separate plasmids. The total initial number of plasmids transfected in a given cell is assumed to follow a log-normal distribution.[26,37] This assumption is because the shape of the protein distribution is known to reflect the shape of the underlying plasmid distribution,[41] and the protein distribution is often observed to be approximately log-normal.[37,38] The conditional distribution of the number of each of the two types of plasmids, given the total number of plasmids, is assumed to be binomial.[26] This is because the plasmids we consider are assumed to be well-mixed, of relatively small and similar sizes, and hence indistinguishable for purposes of co-transfection.[26] In the remainder, we choose values for kinetic parameters such that they span the parameter distributions

calculated from transcriptomics and proteomics data given in Ref. *42*. We select parametric values for the initial plasmid distributions based on the polymerase chain reaction (PCR) findings of Ref. *24*,*40*,*43*. The biochemical reactions in our model are shown below:

$$\begin{aligned}
&\mathrm{D_{tm}} \xrightarrow{K_1} \mathrm{D_{tm}} + \mathrm{M_{tm}} && \mathrm{D_{induced}} \xrightarrow{K_2} \mathrm{D_{induced}} + \mathrm{M_{induced}} \\
&\mathrm{M_{tm}} \xrightarrow{K_3} \mathrm{M_{tm}} + \mathrm{P_{tm}} && \mathrm{M_{induced}} \xrightarrow{K_4} \mathrm{M_{induced}} + \mathrm{P_{induced}} \\
&\mathrm{M_{tm}} \xrightarrow{\Lambda_1} \emptyset && \mathrm{M_{induced}} \xrightarrow{\Lambda_2} \emptyset \\
&\mathrm{P_{tm}} \xrightarrow{\Lambda_3} \emptyset && \mathrm{P_{induced}} \xrightarrow{\Lambda_4} \emptyset,
\end{aligned}$$

where D, M, and P stand for plasmid, mRNA, and protein. Subscript "tm" stands for the transfection marker, and "induced" for the induced gene that is co-transfected. Λ_i ($i = 1-4$) are first order degradation rate constants. Depending on the hypothesis underlying each model, K_i ($i = 1-4$) are defined either as normal first-order rate constants, where $K_1 = k_1 \cdot D_{\mathrm{tm}}$, and K_2, K_3, and K_4 are defined similarly, or as Michaelis-Menten (MM) equations, where a saturated K_1 is defined as $K_{1,\max} \cdot \frac{D_{\mathrm{tm}}}{D_{\mathrm{tm}}+K_{D_{\mathrm{tm}}}}$, and saturated K_2, K_3, and K_4 are defined similarly. $K_{1,\max}$ represents the maximal value of K_1, and $K_{D_{\mathrm{tm}}}$ the half saturation constant. Further details of the models, including plasmid dilution mechanism and length of the simulation, can be found in Supporting Information Section 2.1. Using StochKit and GillesPy, for each fixed set of parameters we simulate this model using the Gillespie method 400,000 times.[29,30,44,45] This is comparable to the number of experimental samples generated in Ref. *26*. After simulation, we divide the simulated data based on the transfection marker into bins of width 0.2, which is comparable to values that are typically chosen in flow cytometry experiments.[26,27,36] We then calculate the geometric mean of the induced protein's concentrations for each bin.

To examine the mechanisms that contribute to the near-constant induced reporter concentrations at low plasmid copy number, and the saturating induced reporter concentrations at high plasmid copy number, we systematically vary individual or pairs of parameters while

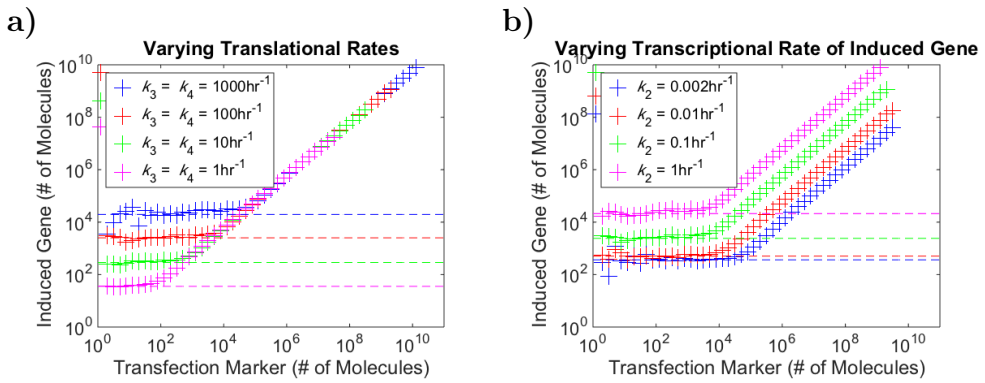


Figure 3: Simulations of our stochastic model suggest that either increasing translation rates (a) or decreasing transcriptional rates (b) can extend the near-constant induced gene levels at low copy plasmid numbers. X-axis and y-axis stand for number of molecules of the transfection marker and the induced protein in each bin. Best fit horizontal lines are drawn for reference. (a) Comparison of models in which the translational rates decrease in order from 1000 to 1 molecule per mRNA per hour. (b) Comparison of models in which the transcriptional rate of D_{induced} increases from 0.002 to 1 molecule per plasmid per hour.

holding the remaining parameters constant. We begin by examining possible mechanisms that lead to near-constant induced reporter concentrations at low plasmid numbers, creating two cohorts of models. In each cohort we assume that K_i are normal first-order rate expressions, i.e., $K_1 = k_1 D_{\text{tm}}$ with K_2, K_3, and K_4 defined similarly. The first cohort varies only the translational rate constants k_3 and k_4, while the second cohort varies only the induced gene's transcriptional rate, k_2. Simulations of the stochastic model demonstrate that either increasing translation rates, or decreasing transcription rates, can lead to the observed constant induced reporter levels at low plasmid copy numbers (Figure 3).

We next investigate mechanisms that may cause protein concentrations to saturate at high plasmid copy numbers. Though the physical mechanism has not been proven, several experimental studies conclude that some steps of the transcription process may saturate in cells expressing large amounts of mRNA.[46,47] It has also been suggested that the cationic liposomes used in transfection inhibit the process of transcription.[48] Hence, it is possible that a high concentration of liposomes (associated with high plasmid copy numbers) is also a mechanism that induces saturation in transcription rates. Motivated by these possible mechanisms, we modify our stochastic model to incorporate saturation of transcriptional

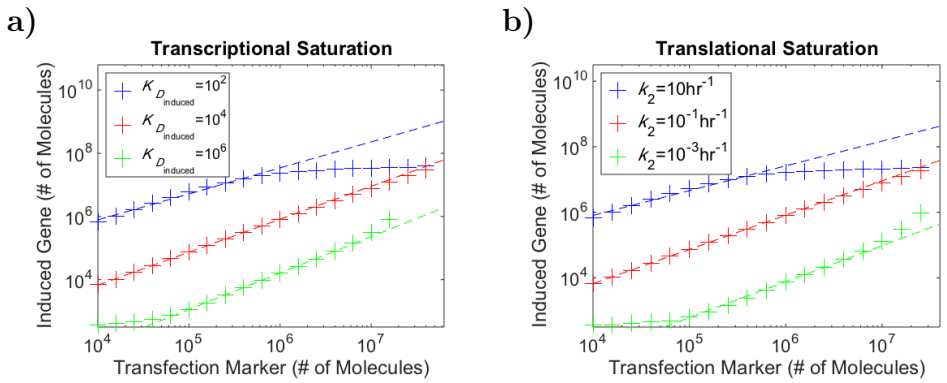


Figure 4: Simulations of our stochastic model suggest that either either saturation of transcriptional kinetics (a) or saturation of translation kinetics (b) can lead to regimes where the induced gene reporter level saturates at high plasmid copy numbers. X-axis and y-axis stand for number of molecules of the transfection marker and the induced protein in each bin. Least squares regression lines are drawn for reference. (a) Comparison of models built under the hypothesis of transcriptional saturation. The half saturation constant $K_{D_{\text{induced}}}$ increases in order from 10^2 to 10^6 molecules, and $K_{D_{\text{tm}}}$ is held fixed at 10^4 molecules. (b) Comparison of models built under the hypothesis of translational saturation. The transcriptional rate of the induced gene decreases in order from 10 to 10^{-3} molecule per plasmid per hour, and the transfection marker transcribes at a constant rate of 10^{-1} molecule per plasmid per hour.

kinetics. We now take the transcription rates, K_1 and K_2, to be given by saturating MM approximations with MM constants, $K_{D_{\text{tm}}}$ and $K_{D_{\text{induced}}}$ (see Supporting Information Section 2.1). Here smaller K_D values correspond to saturation beginning at lower plasmid copy numbers. By systematically varying both K_D values (see Supporting Information Section 2.1) we observe that transcriptional saturation may induce protein saturation when $K_{D_{\text{induced}}} \ll K_{D_{\text{tm}}}$ (see Figure 4(a)). That is, protein levels as a function of the amount of plasmid may saturate if the transcriptional rate of the induced reporter saturates at a lower level of plasmid than that at which the transcriptional rate of the transfection marker saturates.

Finally, we now investigate whether translational saturation can also induce saturation in protein levels at high plasmid copy numbers. Tachibana et al. presented experimental evidence which suggests that protein synthesis saturates when a large amount of mRNA is present.[24] Motivated by this study, we now consider a version of our stochastic model where the transcriptional rates $K_1 = k_1 D_{\text{tm}}$ and $K_2 = k_2 D_{\text{induced}}$ are non-saturating first order reactions as in our first model, but the translation rates K_3 and K_4 are saturating

MM approximations. Since the induced gene and the transfection marker are homologous fluorescent genes, we use the same maximal translation rates and same MM constants in K_3 and K_4 (see Supporting Information Section 2.1). This final version of our model suggests that under the hypothesis of translational saturation, protein reporter saturation can be observed if $k_2 \gg k_1$, i.e. if the induced gene transcribes faster than the transfection marker's gene (see Figure 4(b)).

In summary, we have demonstrated two different physical mechanisms that may induce a near-constant level of the induced gene reporter at low plasmid copy numbers (high translation rates or low transcription rates). We thank a reviewer for pointing out another possible mechanism; that flow cytometry measurements at low plasmid copy numbers are susceptible to experimental noise such as auto-fluorescence, and instrumental limitations. In the absence of experimental noise, but under our modeling assumptions, our stochastic models demonstrate that even with linear production rates a near-constant level of the induced gene reporter will be observed at low plasmid copy number. This arises as the normalized histograms of the plasmid encoding the transfection marker within each of the leftmost bins had relatively constant modes (see Supplementary Figure 7 and Supporting Information Section 2.3 for more details).

Our models also demonstrate two different physical mechanisms that may induce a saturating level of induced gene reporter for high plasmid copy numbers (having the induced gene transcription kinetics saturate at lower plasmid levels than needed for saturation of the transfection marker gene transcription kinetics, or having translational saturation with the induced gene transcribing faster than the transfection marker's gene). Note that the results we have derived do not depend on the precise choice of bin width (see Supplementary Figure 2 in Supporting Information). In Supporting Information Section 2.2 we show that these results persist when considering an alternative model for the initial plasmid distributions within cells. In Supporting Information Section 1.2 we explain why the observed saturation region at high plasmid copy number within the flow cytometry data is unlikely to be due to

experimental noise.

Our analysis poses a challenge to the characterization of circuit behavior in TTMC. The stochastic models demonstrate there are multiple (physical) mechanisms that can explain the observed saturation (constant levels) of the induced gene reporter at high (low) plasmid copy numbers. Due to the complexity of these models it seems unlikely one could fit them, or even select which is most appropriate, from just single-time-point flow cytometry data.

Bin-dependent ODE Model

Though mechanistic details cannot be disentangled from single-time flow cytometry measurements, characterization of modules remains a critical problem to be addressed. This is needed to enable the development of models that can predict the dynamics of circuits/pathways with more components, and, which exhibit more complicated behaviors. To further this goal, we now develop a simple, phenomenological ODE model that can accurately describe single-time transient transfection flow cytometry data. While development of a more physically detailed model would be ideal, as shown in the last subsection it would require additional experimental data to be uniquely determined.

To account for the observed saturation in protein concentration, we propose replacing the traditional Hill-function-based model (see Supporting Information Section 3) with a bin-dependent model. The bin-dependent model divides flow cytometry data into two subsets based on plasmid copy number, i.e., one with and one without saturation.

$$\frac{dI_i}{dt} = \alpha_i \cdot \phi(t) - \lambda \cdot I_i \qquad \phi(t) = \left(\frac{1}{2}\right)^{\lfloor \frac{t}{T} \rfloor}$$

$$\frac{dO_i}{dt} = \begin{cases} \beta \cdot \phi(t) \cdot \left(\dfrac{P_i}{P_1}\right)^f \cdot \left(\dfrac{1-\gamma}{1+\left(\frac{I_i}{d}\right)^h} + \gamma\right) - \lambda \cdot O_i, & \text{if } P_i < P_{i'} \\ \beta \cdot \phi(t) \cdot \left(\dfrac{P_{i'}}{P_1}\right)^f \cdot \left(\dfrac{P_i}{P_{i'}}\right)^g \cdot \dfrac{1-\gamma}{1+\left(\frac{I_i}{d}\right)^h} & \\ \quad + \beta \cdot \phi(t) \cdot \left(\dfrac{P_i}{P_1}\right)^f \cdot \gamma - \lambda \cdot O_i, & \text{if } P_i \geq P_{i'} \end{cases} \tag{1}$$

where I_i and O_i are the concentrations of the input and the output in the i-th bin, and i' is the separating bin. The separating bin is chosen to be the bin at which average concentrations of the co-transfected protein switch from linear growth to saturating growth. α_i, the production rate of the induced protein in the i-th bin, is assumed time-invariant because I_i is induced by a constant concentration of inducer. We do not explicitly characterize the functional form of how α_i depends on the plasmid level as we simply fit a different value of α_i for each bin. $\phi(t)$ captures that the population-average plasmid counts decrease due to cell division over time.[26] T is length of the cell cycle; λ_I and λ_O are dilution/degradation rates of I and O. β is the maximal average production rate of the regulated protein for cells in the 1st bin, i.e, cells that have minimal plasmid counts P_1. P_i is the mid-point of the i-th plasmid count bin. f and g capture the relationship between the concentrations of the transfection marker and the maximal production rates of the output protein for low and high copy numbers, respectively. The bin-dependent model only requires one additional parameter than a standard Hill-function-based model (see Supporting Information Section 3).

We fit the traditional Hill-function-based model (see Supporting Information Section 3) and the bin-dependent model (Equation (1)) to the TAL14, TAL21, and LmrA datasets from Ref. *26* for validation (TAL14, TAL21, and LmrA are names of the repressors).[26] Both models are simulated for 46 hours since an average delay of 25 hours in plasmid expression is expected.[26] Protein loss is assumed to arise purely from dilution, as both the input and output proteins are very stable on the time scale of the experiments.[26] We therefore take $\lambda_I = \lambda_O = \lambda$, and calculate them based on the length of the cell cycle, which spans approximately 20 hours.[26] Davidsohn et al. constructed the circuits using the rtTA and GAL4/UAS system: the input (repressor) is activated by a constitutive rtTA protein and doxycycline, and expression of the output (EYFP), which is inhibited by the input, is driven by a constitutive Gal4 protein.[26] A detailed representation of the circuit structure can be found in Supplementary Figure 1. rtTA and Gal4, which are indispensable for protein acti-

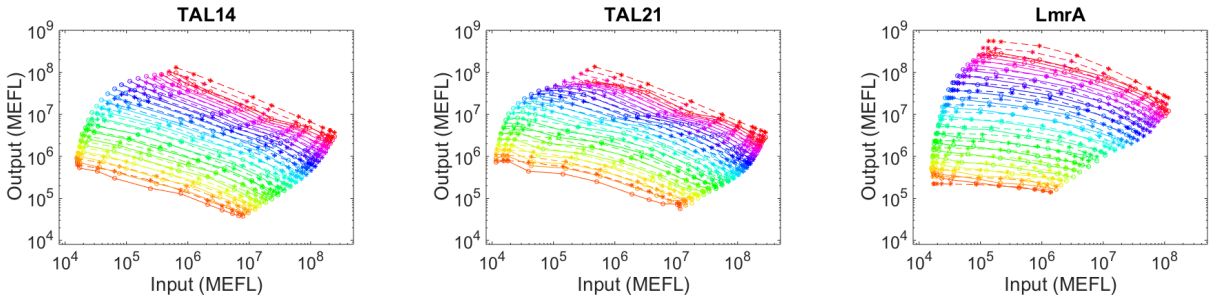


Figure 5: Comparison between experimental data and the traditional Hill-function-based TAL14, TAL21, and LmrA models. Plasmid copy number is shown by color. Solid lines are experimental data, and dashed lines are model fits. The experimental data in the plots are from Ref. *26*.

vation, are both constitutively expressed and are not considered as limiting factors for the production of the input and the output. Omitting rtTA and Gal4 leads to an abstraction of the circuit structure that can be studied by our models, as is shown in Figure 1(a). For the bin-dependent model, the bin that separates flow cytometry data into subsets of fast and slow protein production is chosen to be $10^{7.1}$ MEFL for TAL14 and TAL21, and $10^{7.4}$ MEFL for LmrA since in the dataset, saturation in protein production is observed to the right of 10^{7} MEFL and $10^{7.3}$ MEFL, respectively (Figure 2). Model fitting is implemented via minimizing the mean-squared errors (MSE) between the log of observed and predicted concentrations of the regulated proteins (details of model fitting can be found in Supporting Information Section 4). We log-transform the concentrations to reduce the absolute errors that are often associated with measurements of large protein concentrations on a linear scale.[49] For our specific implementation, we use Matlab's GlobalSearch algorithm to locate the set of parameter values that produce the global minimum error.[50] The optimal parameter fits and the errors in the fit models are shown in Supplementary Tables 2 and 3 in Supporting Information Section 5, and the fit model values versus the experimental values of the fluorescent reporters are shown in Figures 5 and 6. Our results suggest that the bin-dependent model fits the data well for all plasmid copy numbers despite having only one more parameter compared to the Hill-function-based model (Table 1).

We further compare the Hill-function-based model and the bin-dependent model via cross-validation. We conduct a 12-fold cross-validation by randomly dividing the flow cytometry

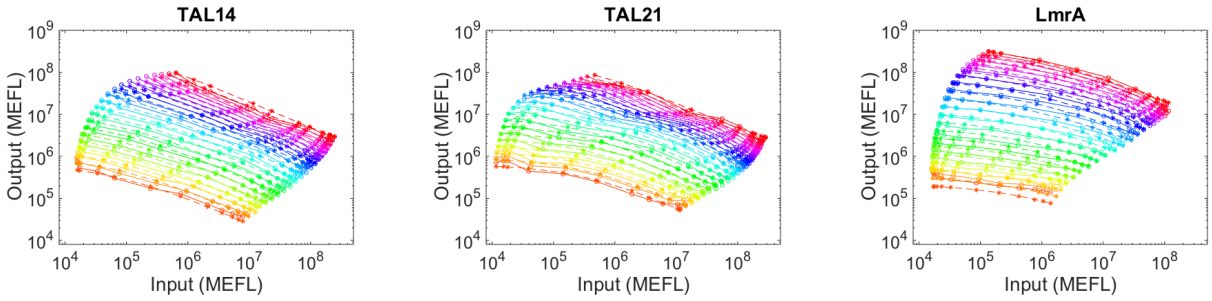


Figure 6: Comparison between experimental data and the bin-dependent TAL14, TAL21, and LmrA models. Plasmid copy number is shown by color. Solid lines are experimental data, and dashed lines are model fits. The experimental data in the plots are from Ref. *26*.

data into 12 subsets of the same size, fitting the models separately on each combination of 11 subsets, and then testing the models on the single subsets that were left out.[51] The fitting errors and the testing errors are then averaged over the 12 combinations of subsets. Our results suggest that both the fitting errors and the testing errors of the bin-dependent models are 1.5 - 2 times better than those of the Hill-function-based models (Tables 2 and 3). The bin-dependent model shows a less significant improvement for LmrA than for TAL14 and TAL21. A possible explanation is that for LmrA, the saturation effect is observed in six bins to the right of $10^{7.3}$ MEFL rather than in nine bins to the right of 10^{7} MEFL. For each repressor, we choose the model that produces the least testing error among 12 cross-validated models to be the best model. We evaluate the best models for each plasmid copy number. The results indicate that the bin-dependent models produce not only lower but also more consistent errors across all bins (Figure 7). The errors of the Hill-function-based models get large near 10^{7} MEFL and $10^{7.8}$ MEFL for all repressors. This signals that there are patterns in the data that are not explained by the Hill-function-based models.[52] The bin-dependent model produces larger errors for LmrA than for TALER repressors because there are slight indications of a near-constant region at low plasmid numbers for LmrA (Figure 2). In summary, we find that the bin-dependent model consistently provides significantly better fits to the experimental data than the Hill-function-based model.

Note, for high-plasmid-count subsets, our bin-dependent model assumes the log of the maximal protein production rate is approximated as a linear function of the log of the

Table 1: MSE of the models.

Goodness of fit		
Repressor	Hill-function-based	bin-dependent
TAL14	0.013	0.004
TAL21	0.015	0.005
LmrA	0.020	0.009

Table 2: Averaged fitting errors of the models within the 12-fold cross-validation.

Fitting Errors[a]		
Repressor	Hill-function-based	bin-dependent
TAL14	0.013	0.006
TAL21	0.017	0.009
LmrA	0.018	0.013

[a] See Supporting Information Section 4 for the definition of fitting errors.

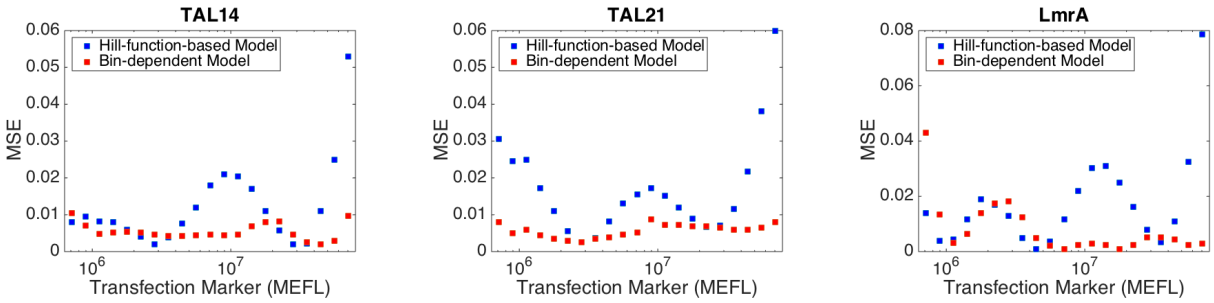


Figure 7: Testing errors of the best cross-validated models within each bin.

transfection marker. Although the relationship is arguably better fit by other functions, our assumption leads to a model with a good fit across the entire dataset, while only requiring one additional parameter.

Modular Composition

To validate the predictive power of the bin-dependent model, we develop models for the six two-repressor cascades and three of the feed-forward circuits shown in Ref. *26* (for which we were given the experimental data from Ref. *26*). The exact structure of the cascades and the feed-forward circuits can be found in Figures 3(A) and 5(A) of Ref. *26* or Supporting Information Supplementary Figures 1(b) and 1(c), with Figures 8(a) and 10(a) providing abstractions that highlight the key parts of the circuits. A two-repressor cascade can be

Table 3: Averaged testing errors of the models within the 12-fold cross-validation.

Testing Errors[b]		
Repressor	Hill-function-based	bin-dependent
TAL14	0.014	0.007
TAL21	0.017	0.008
LmrA	0.019	0.013

[b] See Supporting Information Section 4 for the definition of testing errors.

decoupled into two modules, with the output of the first module acting as the input of the second module (Figure 8(a)). Similarly, a feed-forward circuit can be decoupled into three modules (Figure 10(a)). The bin-dependent models for cascades and feed-forward circuits are constructed, and their agreement with experimental measurements are compared with that of the Hill-function-based and the EQuIP models developed in Ref. *26*. Specifically, we compare simulations of the circuit models to experimental data by measuring the differences between simulated and observed concentrations of EYFP 72 hours post transfection (experimental data from Ref. *26*). Full details of the experimental protocol can be found in Ref. *26*. The equations and parameters for the bin-dependent models can be found in Supporting Information Section 6 and Supporting Information Section 8. The bin-dependent circuit models are developed by composing together the individual module models that were *individually fit* in the previous section. We *do not re-fit* the equations for each model to data for the complete two-module cascades or three-module feed-forward circuits. In this way we can assess how well models fit to individual modules can predict circuit behavior when composed together. To offer a comparable study to Ref. *26*, we use the parameters Davidsohn et al. fit for the Hill-function-based models for cascades.[26] Hill-function-based models for feed-forward circuits were not studied in Ref. *26*. We therefore construct Hill-function models of feed-forward circuits by composing the parameterized Hill-function models of individual modules developed in the previous section.

Like most biological data, calibrated flow cytometry is subject to batch effects. Parameters in the models of the modules need to be rescaled so that they are brought to the same scale before the models are connected into a circuit. Rescaling is a two-step process, where

systematic variation between modules is first removed to facilitate modular composition, and then reincorporated in the model to enable a fair comparison between model predictions and experimental data. The rescaling method we used can be found in Supporting Information Section 6. Davidsohn et al. determined the scaling factors among batches directly from experimental data by comparing the means and the tightness of the data of different batches for all the modules and cascades (note this calculation does not rely on EQuIP).[26] Values of the scaling factors for each input protein I, output protein O, and transfection marker can be found in Section 12 of the Supporting Information of Ref. *26*. We use these scaling factors to rescale the parameters of the bin-dependent models since rescaling in our context is first-order linear compensation,[26] i.e. there is no difference between rescaling the parameters and fitting the parameters to rescaled data. To understand the effect of cross-batch compensation on model predictions, we also rescale the parameters of the Hill-function models provided in Ref. *26* by the same scaling factors. Since scaling factors for both modules and cascades are provided in Ref. *26*, we perform both steps of rescaling amid construction of cascade models. The equations and parameters for the bin-dependent model and the rescaled Hill-function-based models for the cascades can be found in Supporting Information Section 6. For feed-forward circuits, we did not have experimental data with which to calculate scaling factors, and so we only performed parameter rescaling at the modular level. The equations and parameters for the feed-forward circuits can be found in Supporting Information Section 8.

The agreement between experimental measurements and model predictions for the six cascades is illustrated in Figure 9 and the figures of Supporting Information Section 7. For all six cascades, the bin-dependent model is able to capture the positive association between the input and the output (Figure 9). It also captures the buffer-like behavior of the cascades, i.e., the dynamic range of the output is narrower compared to that of the input due to low cooperativity of the regulatory modules (Figure 9 and Supporting Information Section 7).[53]

To further investigate how well our composed circuit models fit the experimental data,

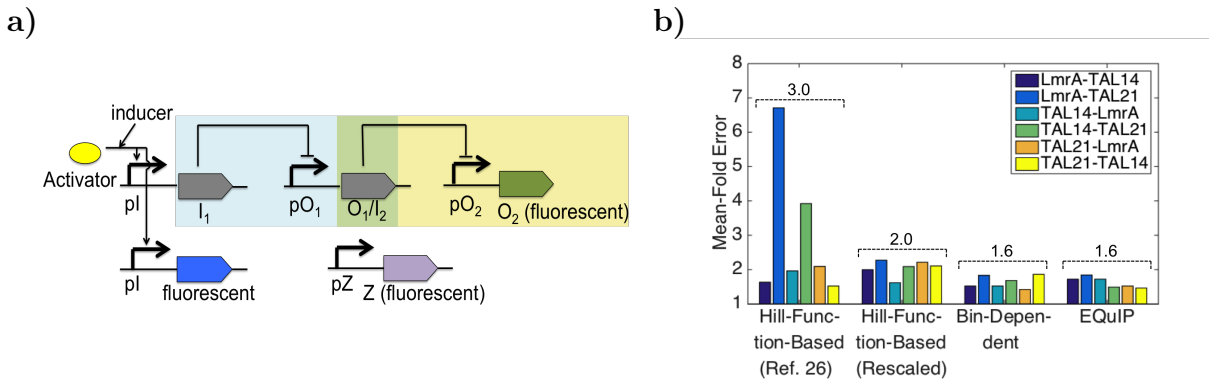


Figure 8: (a) Abstraction of the structure of the cascade. I_1 inhibits I_2/O_1, which further inhibits O_2. The expression of I_2/O_1 and O_2 is driven by a constitutive Gal4 protein and is omitted from the plot. The overlapping component of the modules is shown in the blended color. (b) Comparison of the mean-fold errors of the Hill-function-based models, with and without rescaling,[26] the bin-dependent models, and EQuIP[26] for each cascade. The experimental data the models are validated against are from Ref. *26*. Numbers on top of the dotted lines represent the average mean-fold errors of six cascades.

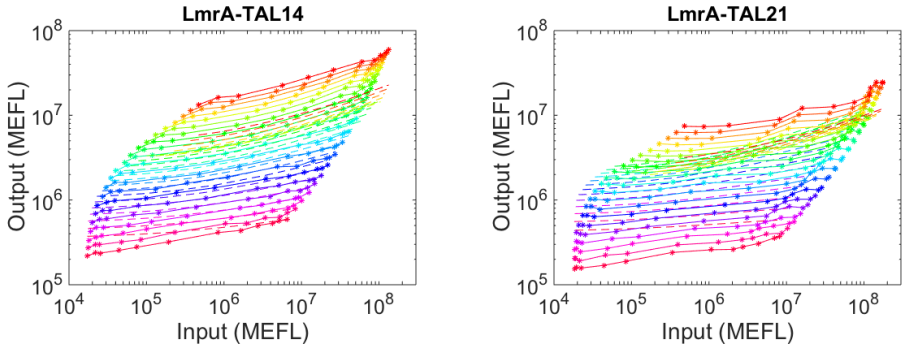


Figure 9: Comparison between experimental data and predictions made by the bin-dependent models for LmrA-TAL14 and LmrA-TAL21 cascades. Plasmid copy number is shown by color. Solid lines are experimental data, and dashed lines are model fits. Experimental data in the plots are from Ref. *26*.

we examined the average mean fold error, defined as the average over all six cascades of the mean-fold errors over all induction levels of each individual cascade (see Supporting Information Section 4 for details and formulas). The rescaled bin-dependent model is found to outperform the Hill-function-based model presented in Ref. *26*, with an average mean-fold error of 1.6 fold for the former vs 3.0 fold for the latter. The 1.6 fold average error of the bin-dependent model also outperforms the average error of the Hill-function model with rescaling, which was found to be 2.0 fold (Figure 8(b)). This indicates that inconsistent scales due to batch effects contribute significantly to the magnitude and the inconsistency

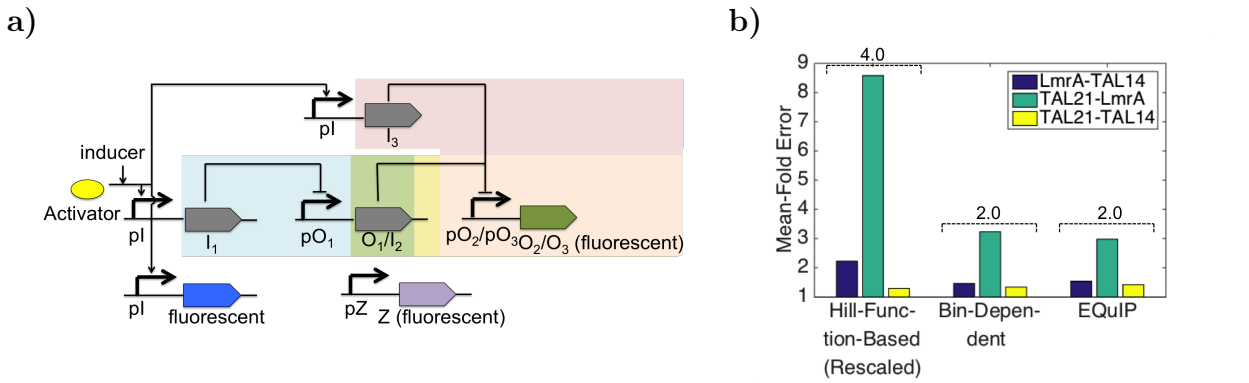


Figure 10: (a) Abstraction of the structure of the feed-forward circuit. I_1 inhibits I_2/O_1, which further inhibits O_2. I_3 inhibits O_3. The expression of I_2/O_1 and O_2/O_3 is driven by a constitutive Gal4 protein and is omitted from the plot. The overlapping components of the modules are shown in the blended colors. (b) Comparison of the mean-fold errors of the rescaled Hill-function-based models, the bin-dependent models, and EQuIP[26] for each feed-forward circuit. The experimental data the models are validated against are from Ref. *26*. Numbers on top of the dotted lines represent the average mean-fold errors of three feed-forward circuits.

of the errors. In addition, the rescaled bin-dependent model also produces smaller mean-fold errors than the rescaled Hill-function model for all individual cascades (Figure 8(b)). In Ref. *54*, we examined a different rescaled Hill-function model, based on composing the Hill-function models we parameterized for individual modules in the preceding subsection. For this rescaled Hill-function model we observed an average mean fold error of 1.8.[54] As we fit the Hill-function model using a different optimization routine than used in Ref. *26*, this illustrates that the parameter estimation procedure can also influence the relative accuracy of different models. Finally, we note that the accuracy of the bin-dependent model varies relative to EQuIP, achieving a smaller mean-fold error for some cascades and larger error others (see Figure 8(b)). The average over all six cascades is the same as EQuIP (1.6), which is considered high accuracy based on results reported in the literature.[26,27,34,55,56]

Besides cascades, the bin-dependent model also facilitates relatively accurate predictions for feed-forward circuits. The agreement between experimental measurements and model predictions for the three feed-forward circuits is illustrated in the figures of Supporting Information Section 9. The average error over all three feed-forward circuits is the same as EQuIP (2.0) and is much lower than the rescaled Hill-function-based model (4.0) (Figure

10(b)). The bin-dependent model captures the qualitative behavior of the circuit – the output is weakly affected by a change in the input at low inducer levels due to two opposing regulations: $I_1 -| \dashv O_1 \dashv O_2$ and $I_3 \dashv O_3$, and negatively associated with the input at high inducer levels as inhibition becomes the dominant force. The relatively large error for the TAL21-LmrA circuit is likely to be batch-specific, as measurements of the output are below 10^4 MEFL at low plasmid copy number (see Supplementary Figure 10). Note, such low levels of MEFL are not observed in any of the other circuit datasets.

Despite the relatively high accuracy of the bin-dependent model, we note that the simplicity of its representation of the gene expression process may in some contexts sacrifice accuracy. Figure 9 shows an under-prediction for two out of six cascades, the cause for which may be attributed to the non-negligible amount of time over which transcription and translation take place. The time lag between expression of I_2/O_1 and O_2 may be better captured by delay differential equations (DDE).

Conclusions

We have developed a bin-dependent ODE model that describes regulatory mechanisms via the use of standard Hill function type terms, while offering comparable accuracy to the EQuIP model of Ref. *26*. Parameterized, bin-dependent models of individual modules should be relatively straightforward to integrate as subcomponents within larger existing ODE and DDE models of circuits. Moreover, it should also be relatively straightforward to modify a parameterized bin-dependent model to incorporate *additional*, previously-characterized regulatory components (i.e. for studying promoters co-regulated by multiple transcription factors). In this way we expect that bin-dependent models for individual modules should be able to be composed with a variety of existing, well-characterized differential equation models that describe components of synthetic and systems biology networks.

Another benefit to the bin-dependent-model-based approach is that it is fairly robust to

sampling noise in experimental data. The input-output datasets, which the ODE models are fit to, comprise the geometric means of measured protein concentrations within each bin. These data points may not be well separated, and hence appear noisy, when using sparse flow cytometry datasets. The model fitting step helps overcome this sampling noise by using deterministic ODEs based on widely-used biochemical relationships (such as Hill-functions).

The bin-dependent model presented here establishes a framework for characterizing fundamental synthetic constructs and predicting circuit behaviors quantitatively in TTMC. As we demonstrated with the stochastic model, there are different mechanisms that may contribute to saturation in protein production, a common phenomenon in TTMC. The value of the bin-dependent model lies in both its easy integrability with other differential equation models, and in its ability to describe the saturation effect in flow cytometry data accurately without specifying precise mechanistic details for how saturation occurs. The method presented here should be applicable to similar flow cytometry datasets, allowing the possibility to construct a well-characterized library of in silico models for regulatory switches. The quantitative parameters of such regulatory switches could then be used in constructing new predictive models for the behaviors of more complicated circuits. Our work represents one more step towards building a systematic workflow that can guide circuit design in TTMC.

Associated Content

Supporting Information Available

Experimental details, two-stage gene expression models, Hill-function-based models, details of model fitting, optimal parameter fits of ODE models, models for cascades, cascade predictions, models for feed-forward circuits, and feed-forward circuit predictions. If accepted, This material is available free of charge via the Internet at `http://pubs.acs.org/`.

Author Information

Corresponding Author

*E-mail: dawang@bu.edu.

Author Contributions

J.W. analyzed data, developed and conducted computational analysis, and wrote the manuscript. S.A.I and C. B. helped develop the computation framework and models, and wrote the manuscript.

Acknowledgement

The authors thank Jacob Beal, Prof. Chris Myers, Prof. Daniel Segre, Brian Teague, and Changzhe Tian for helpful discussions and constructive feedbacks. We thank Jacob Beal for suggesting the initial condition for plasmid copy numbers we used in the subsection "Protein Concentration vs Plasmid Copy Number" and Supporting Information Section 2.1 based on his observations from the studies in Davidsohn et al. (2015). We also thank Jacob Beal and Brian Teague for providing us the data and the scripts from Ref. *26*. This work was supported by the National Science Foundation under Grant No. `CNS-1446607` and Grant No. `CBET-0939511`. SAI was partially supported by National Science Foundation award `DMS-1255408`.

Abbreviations

(TTMC), transiently transfected mammalian cells; (ODE), ordinary differential equations; (PCR), polymerase chain reaction; (MEFL), Molecules of Equivalent Fluorescein; (MM), Michaelis-Menten; (DDE), delay differential equations

References

1. Dalton, A., and Barton, W. (2014) Over-expression of secreted proteins from mammalian cell lines. *Protein Sci. 23*, 517–525.

2. Khan, K. (2013) Gene expression in mammalian cells and its applications. *Adv. Pharm. Bull. 3*, 257–263.

3. Vink, T., Oudshoorn-Dickmann, M., Roza, M., Reitsma, J.-J., and de Jong, R. N. (2014) A simple, robust and highly efficient transient expression system for producing antibodies. *Methods 65*, 5–10.

4. Khalil, A. S., and Collins, J. J. (2010) Synthetic biology: applications come of age. *Nat. Rev. Genet. 11*, 367–379.

5. Schenborn, E. T., and Goiffon, V. (2000) DEAE-dextran transfection of mammalian cultured cells. *Methods Mol. Biol. 130*, 147–153.

6. Smith, C. Stable vs. transient transfection of eukaryotic cells. `http://www.biocompare.com/Editorial-Articles/126324-Transfection/`.

7. Recillas-Targa, F. (2006) Multiple strategies for gene transfer, expression, knockdown, and chromatin influence in mammalian cell lines and transgenic animals. *Mol. Biotechnol. 34*, 337–354.

8. Kis, Z., Pereira, H. S., Homma, T., Pedrigi, R. M., and Krams, R. (2015) Mammalian synthetic biology: emerging medical applications. *J. R. Soc. Interface 12*, 20141000.

9. Kim, T., and Eberwine, J. (2010) Mammalian cell transfection: the present and the future. *Anal. Bioanal. Chem. 397*, 3173–3178.

10. Schaumberg, K. A., Antunes, M. S., Kassaw, T. K., Xu, W., Zalewski, C. S., Medford, J. I., and Prasad, A. (2016) Quantitative characterization of genetic parts and circuits for plant synthetic biology. *Nat. Methods 13*, 94–100.

11. Wurm, F. M. (2004) Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat. Biotechnol. 22*, 1393–1398.

12. Rekhi, R., and Qutub, A. (2013) Systems approaches for synthetic biology: a pathway toward mammalian design. *Front. Physiol. 4*, 285.

13. Mathur, M., Xiang, J. S., and Smolke, C. D. (2017) Mammalian synthetic biology for studying the cell. *J. Cell Biol. 216*, 73–82.

14. (2014) Synthetic biology: back to the basics. *Nat. Methods 11*, 463.

15. Gyorgy, A., and Del Vecchio, D. (2014) Modular composition of gene transcription networks. *PLOS Comput. Biol. 10*, e1003486.

16. Del Vecchio, D., and Sontag, E. D. Dynamics and control of synthetic bio-molecular networks. Proc. ACC. New York, 2007.

17. Del Vecchio, D., Qian, Y., and Dy, A. (2016) Control theory meets synthetic biology. *J. R. Soc. Interface 13*, 20160380.

18. Sivakumar, H., and Hespanha, J. Towards modularity in biological networks while avoiding retroactivity. Proc. ACC. 2013.

19. Brophy, J. A. N., and Voigt, C. A. (2014) Principles of genetic circuit design. *Nat. Methods 11*, 508–520.

20. Densmore, D., and Hassoun, S. (2012) Guest Editors' Introduction: Synthetic Biology. *IEEE Des. Test Comput. 29*, 5–6.

21. Basu, S., Gerchman, Y., Collins, C., Arnold, F., and Weiss, R. (2005) A synthetic multicellular system for programmed pattern formation. *Nature 434*, 1130–1134.

22. Gardner, T., Cantor, C., and Collins, J. (2000) Construction of a genetic toggle switch in Escherichia coli. *Nature 403*, 339–342.

23. Ellis, T., Wang, X., and Collins, J. (2009) Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat. Biotechnol. 27*, 465–471.

24. Tachibana, R., Harashima, H., Ide, N., Ukitsu, S., Ohta, Y., Suzuki, N., Kikuchi, H., Shinohara, Y., and Kiwada, H. (2002) Quantitative analysis of correlation between number of nuclear plasmids and gene expression activity after transfection with cationic liposomes. *Pharm. Res. 19*, 377–381.

25. Glover, D. J., Leyton, D. L., Moseley, G. W., and Jans, D. A. (2010) The efficiency of nuclear plasmid DNA delivery is a critical determinant of transgene expression at the single cell level. *J. Gene Med. 12*, 77–85.

26. Davidsohn, N., Beal, J., Kiani, S., Adler, A., Yaman, F., Li, Y., Xie, Z., and Weiss, R. (2015) Accurate predictions of genetic circuit behavior from part characterization and modular composition. *ACS Synth. Biol. 4*, 673–681.

27. Davidsohn, N. Foundational platform for mammalian synthetic biology. Ph.D. thesis, Massachusetts Institute of Technology, 2013.

28. Stanton, B., Siciliano, V., Ghodasara, A., Wroblewska, L., Clancy, K., Trefzer, A., Chestnut, J., Weiss, R., and Voigt, C. (2014) Systematic transfer of prokaryotic sensors and circuits to mammalian cells. *ACS Synth. Biol. 3*, 880–891.

29. Thattai, M., and van Oudenaarden, A. (2004) Stochastic gene expression in fluctuating environments. *Genetics 167*, 523–530.

30. Gillespie, D. T. (1977) Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem. A 81*, 2340–2361.

31. Kærn, M., Blake, W., and Collins, J. (2003) The engineering of gene regulatory networks. *Annu. Rev. Biomed. Eng. 5*, 179–206.

32. Chen, H., and Xia, H. (2011) The research and application of Tet-induced regulatory systems. *Chem. of Life 31*, 285.

33. Assur, Z., Hendrickson, W. A., and Mancia, F. (2012) Tools for coproducing multiple proteins in mammalian cells. *Methods Mol. Biol. 801*, 173–187.

34. Beal, J. (2015) Bridging the gap: a roadmap to breaking the biological design barrier. *Front. Bioeng. Biotechnol. 2*, 87.

35. *Measuring Molecules of Equivalent Fluorescein (mefl), pe (mepe), and rpe-cy5 (mepcy) Using Sphero Rainbow Calibration Particles*; 2001.

36. Siciliano, V., DiAndreth, B., Monel, B., Beal, J., Huh, J., Clayton, K. L., Wroblewska, L., McKeon, A., Walker, B. D., and Weiss, R. (2018) Engineering modular intracellular protein sensor-actuator devices. *Nat. Commun. 9*, 1881.

37. Beal, J. (2017) Biochemical complexity drives log-normal variation in genetic expression. *Engineering Biology 1*, 55–60.

38. Hattis, D., and Burmaster, D. E. (2006) Assessment of variability and uncertainty distributions for practical risk analyses. *Risk Anal. 14*, 713–730.

39. Alon, U. *An Introduction to Systems Biology - Design Principles of Biological Circuits*; Chapman and Hall, 2007.

40. Cohen, R. N., van der Aa, M., Macaraeg, N., Lee, A., and Jr., F. C. S. (2009) Quantification of plasmid DNA copies in the nucleus after lipoplex and polyplex transfection. *J. Control. Release 135*, 166–174.

41. Tal, S., and Paulsson, J. (2012) Evaluating quantitative methods for measuring plasmid copy numbers in single cells. *Plasmid 67*, 167–173.

42. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature 473*, 337–342.

43. B. James, M., and Giorgio, T. (2000) Nuclear-associated plasmid, but not cell-associated plasmid, Is correlated with transgene expression in cultured mammalian cells. *Mol. Ther. 1*, 339–346.

44. Abel, J. H., Drawert, B., Hellander, A., and Petzold, L. R. (2016) GillesPy: a Python package for stochastic model building and simulation. *IEEE Life Sci. Lett. 2*, 35–38.

45. Sanft, K. R., Wu, S., Roh, M., Fu, J., Lim, R. K., and Petzold, L. R. (2011) StochKit2: software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics 27*, 2457–2458.

46. Takahashi, Y., Nishikawa, M., Takiguchi, N., Suehara, T., and Takakura, Y. (2011) Saturation of transgene protein synthesis from mRNA in cells producing a large number of transgene mRNA. *Biotechnol. Bioeng. 108*, 2380–2389.

47. Hama, S., Akita, H., Ito, R., Mizuguchi, H., Hayakawa, T., and Harashima, H. (2006) Quantitative comparison of intracellular trafficking and nuclear transcription between adenoviral and lipoplex systems. *Mol. Ther. 13*, 786–794.

48. Tachibana, R., Harashima, H., Shinohara, Y., and Kiwada, H. (2001) Quantitative studies on the nuclear transport of plasmid DNA and gene expression employing nonviral vectors. *Adv. Drug Deliv. Rev. 52*, 219–226.

49. Braun, D., Basu, S., and Weiss, R. Parameter estimation for two synthetic gene networks: a case study. Proc. IEEE. Int. Conf. Acoust. Speech Signal Process. 2005; pp 769–772.

50. Ugray, Z., Lason, L., Plummer, J., Glover, F., Kelly, J., and Marti, R. (2007) Scatter

search and local NLP solvers: a multistart framework for global optimization. *INFORMS J. Comput. 19*, 328–340.

51. Geisser, S. *Predictive Inference: an Introduction*; Chapman and Hall: New York, NY, 1993.

52. Martin, J., Daffos, D., de Adana, R., and Asuero, A. G. In *Fitting Models to Data: Residual Analysis, a Fitting Models to Data: Residual Analysis, a Primer, Uncertainty Quantification and Model Calibration*; Hessling, D. J. P., Ed.; InTech, 2017.

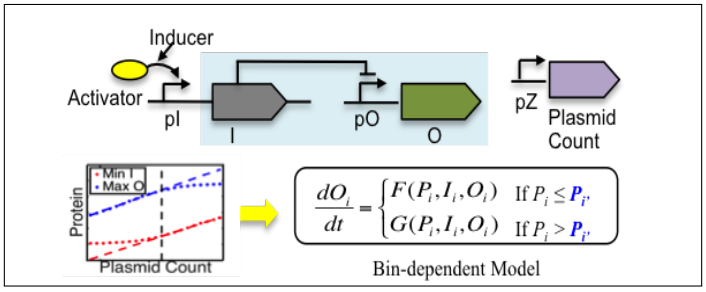
53. Ferrell, J. E., and Ha, S. H. (2014) Ultrasensitivity part III: cascades, bistable switches, and oscillators. *Trends Biochem. Sci. 39*, 612–618.

54. Wang, J., Isaacson, S. A., and Belta, C. Predictions of Genetic Circuit Behaviors Based on Modular Composition in Transiently Transfected Mammalian Cells. 2018 IEEE Life Sciences Conference (LSC). 2018; pp 85–88.

55. Beal, J., Wagner, T. E., Kitada, T., Azizgolshani, O., Parker, J. M., Densmore, D., and Weiss, R. (2015) Model-driven engineering of gene expression from RNA replicons. *ACS Synth. Biol. 4*, 48–56.

56. Wagner, T. E. Engineering a regulatory framework for synthetic self-amplifying RNA circuits. Ph.D. thesis, Boston University, 2017.

Graphical TOC Entry

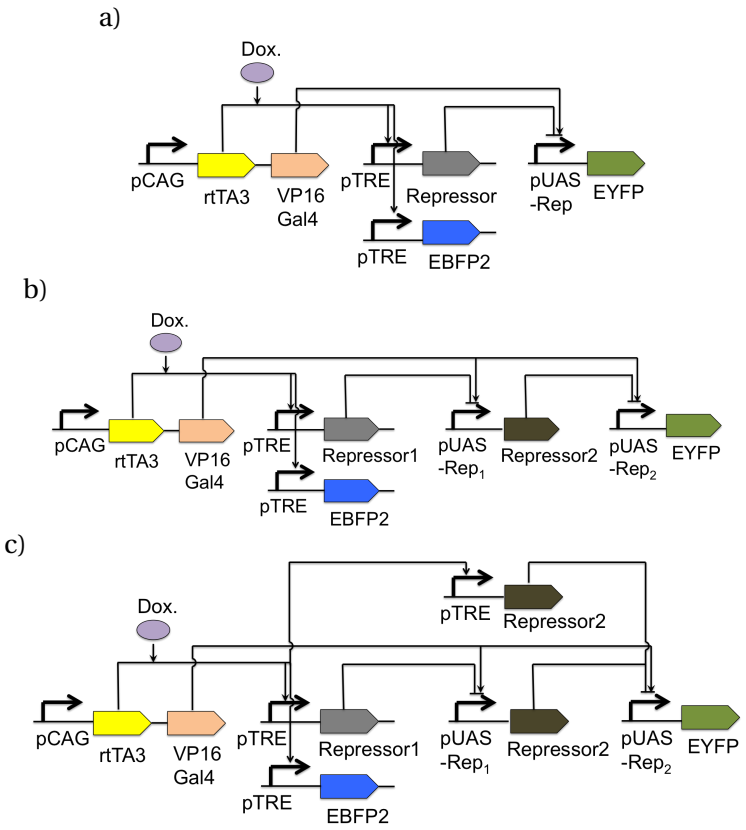


Supporting Information

Modeling Genetic Circuit Behavior in Transiently Transfected Mammalian Cells

Junmin Wang, Samuel A. Isaacson, Calin Belta

Contents



Supplementary Figure 1: Detailed representations of (a) an inducible switch network, (b) a cascade, and (c) a feed-forward circuit controlled by doxycycline based on Figures 2(A), 3(A), and 5(A) of Davidsohn et al. (2015). The transcriptional repressors can be TAL14, TAL21, or LmrA. Expressions of the repressors (TAL14, TAL21, or LmrA) and EYFP are driven by constitutive rtTA and Gal4 proteins, respectively. rtTA and Gal4, which are required for protein activation, are both constitutively expressed and are not considered as limiting factors for the production of the repressors and EYFP.

1 Experimental Details

1.1 Summary of Experimental Details from Davidsohn et al. (2015)

A transcriptional regulatory switch is constructed by connecting each of the three repressors with promoter pUAS-Rep, which controls the expression of a fluorescent gene, EYFP (Supplementary Figure 1(a)). The strength of repression is modulated by inducing the switch at twelve dosages of doxycycline (Dox), and is indicated by the reporter gene EBFP2. Another fluorescent gene, mKate, is a constitutively expressed gene that serves as a transfection marker. All three fluorescent proteins are highly stable, and loss of protein concentration is assumed to be due to dilution (Davidsohn et al. (2015)). A cascade (Supplementary Figure 1(b)) is constructed via the connection of two switches, where the output protein of the first switch acts as the input of the second switch. The plasmids encoding the second repressor of each cascade are transfected at one third the concentration of the plasmids encoding the

first repressor (Davidsohn et al. (2015)).

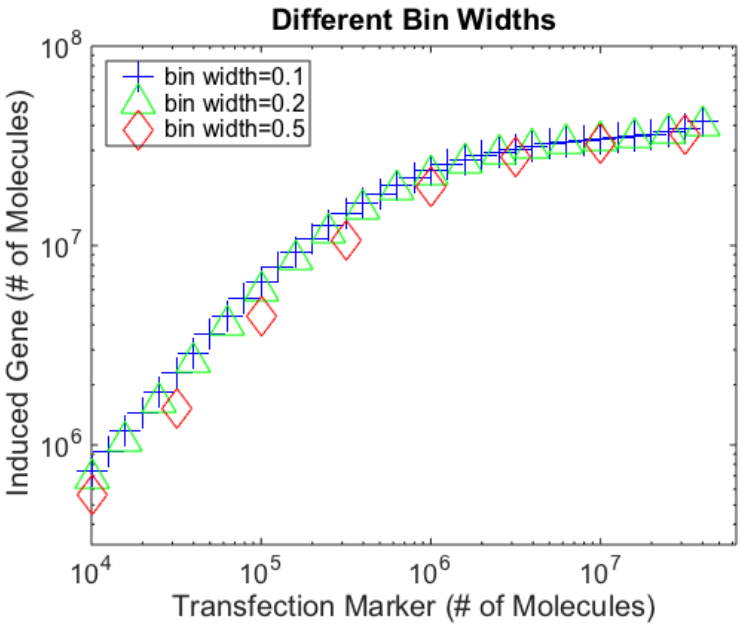
Concentrations of all fluorescent proteins are measured for every single cell by a flow cytometer 72 hours post transfection (Davidsohn et al. (2015)). These data are then standardized into MEFL units and segmented by concentrations of the mKate protein into bins of width 0.1 on a log scale. Because bi-modality observed in the concentrations of the mKate protein is believed to be caused by whether individual cells get transfected, only cells with concentrations of mKate centering around the larger mode, ranging from $10^{5.8}$ to $10^{7.9}$ (unit: MEFL), are used for modeling as in Davidsohn et al. (2015). For data that lie in this range, geometric means of concentrations of the EBFP2 protein and the EYFP protein are calculated within each bin. As plasmids are only expressed after entering the nucleus during mitosis, a delay in plasmid expression is expected (Davidsohn et al. (2015)). The cells are asynchronous: active expression is observed in a fraction of cells 15 hours post transfection, but an average initial delay is estimated to be 25 hours for the entire population (Davidsohn et al. (2015)).

In our models for mean concentrations, we are focused on predicting the average behavior across the cell population and hence ignore this variability. More sophisticated stochastic models could be developed to explicitly account for the variability in the initiation of expression if needed. Length of the cell cycle is measured to be approximately 20 hours (Davidsohn et al. (2015)). All values mentioned above can be confirmed in Davidsohn et al. (2015).

More details of the experiment, including cell culturing, transfection, flow cytometry, and cloning, can also be found in Davidsohn et al. (2015).

1.2 Possibility of Experimental Noise as the Cause of Saturation

We note that the special regions at low and high plasmid numbers (Figure 2) could be speculated to arise from the limited detection range of the flow cytometer. Data in Davidsohn et al. (2015) suggest that the upper detection limit is at least $10^{9.2}$ MEFL (Supplementary Figure 24(a) of Davidsohn et al. (2015)). The possibility of a detection limit can then be ruled out at high plasmid numbers for two reasons. First, the induced and the regulated proteins saturate near 10^8 and 10^7 MEFL, respectively (Figure 2). Near 10^8 and 10^7 MEFL, the geometric standard deviations of (MEFL) concentrations of the induced protein and the regulated protein are between 2 and 2.5. Protein concentrations within each bin are approximately lognormal distributed (Beal (2017)), which means 95% of the cells are within two geometric standard deviations from the geometric means, which is less than $10^{9.2}$ MEFL. In other words, there are fewer than 2.5% of the cells whose fluorescence intensity exceeds $10^{9.2}$ MEFL. Hence, the upper limit of the detection range at $10^{9.2}$ does not have substantial effects on the reported values of our data. Second, saturations due to instrument range often cause protein histograms to have an abrupt cut-off shape, i.e., measurements exceeding the upper detection limit would all gather near a single value (see Supplementary Figure 16(b), Supplementary Figure17(b), and Supplementary Figure18(b) of Davidsohn et al. (2015)). At low plasmid numbers, autofluorescence is a major obstacle limiting the detection sensitivity (Brahme (2014)). Despite autofluorescence corrections, data towards the lower end may be susceptible to experimental noise. Our stochastic models provide an alternative approach to studying these systems with low numbers of molecules. The simulations suggest the possibility of near-constant average protein levels in minimally transfected cells when flow cytometry



Supplementary Figure 2: Simulations of a transcriptional saturation model. X-axis stands for the mid point of each bin, and y-axis number of molecules of the induced protein in each bin. Bin width is chosen to be 0.1, 0.2, and 0.5. Notice, the saturating effect and the general curve are independent of bin size.

measurement noise is removed.

2 Two-stage Stochastic Gene Expression Models

2.1 Model Details

Details regarding the two-stage stochastic gene expression models can be found in this section.

The induced gene and the transfection marker are encoded on separate plasmids. Gene expression is modeled as a two-stage process consisting of transcription and translation. Length of the simulation is 50 hours. Cell division takes place every 20 hours, and plasmids are binomially partitioned in daughter cells upon cell division. The initial cell cycle position for a cell is sampled randomly from the uniform distribution unif(0,20). The reaction rates can be expressed as follows:

$$K_1 = k_1 \cdot D_{\text{tm}}, \qquad K_2 = k_2 \cdot D_{\text{induced}},$$
$$K_3 = k_3 \cdot M_{\text{tm}}, \qquad K_4 = k_4 \cdot M_{\text{induced}},$$
$$\Lambda_1 = \lambda_1 \cdot M_{\text{tm}}, \qquad \Lambda_2 = \lambda_2 \cdot M_{\text{induced}},$$
$$\Lambda_3 = \lambda_3 \cdot P_{\text{tm}}, \qquad \Lambda_4 = \lambda_4 \cdot P_{\text{induced}},$$

Parameter Values								
Figure #	k_1	k_2	k_3	k_4	λ_1	λ_2	λ_3	λ_4
Figure 3(a) and Supplementary Figure 3	0.1	0.1	1000	1000	0.01	0.01	0.01	0.01
	0.1	0.1	100	100	0.01	0.01	0.01	0.01
	0.1	0.1	10	10	0.01	0.01	0.01	0.01
	0.1	0.1	1	1	0.01	0.01	0.01	0.01
Figure 3(b) and Supplementary Figure 4	0.1	0.002	100	100	0.01	0.01	0.01	0.01
	0.1	0.01	100	100	0.01	0.01	0.01	0.01
	0.1	0.1	100	100	0.01	0.01	0.01	0.01
	0.1	1	100	100	0.01	0.01	0.01	0.01
Figure 4(a) and Supplementary Figure 5	NA	NA	100	100	0.01	0.01	0.01	0.01
	NA	NA	100	100	0.01	0.01	0.01	0.01
	NA	NA	100	100	0.01	0.01	0.01	0.01
Figure 4(b) and Supplementary Figure 6	0.1	10	NA	NA	0.01	0.01	0.01	0.01
	0.1	0.1	NA	NA	0.01	0.01	0.01	0.01
	0.1	0.001	NA	NA	0.01	0.01	0.01	0.01

Supplementary Table 1: Parameter values for the two-stage models. k_1 and k_2 have the units of # of molecules per plasmid per hour. k_3 and k_4 have the units of # of molecules per mRNA per hour. λ_1, λ_2, λ_3, and λ_4 have the units of reciprocal hours. In models corresponding to Figure 4(a) of the main text and Supplementary Figure 5, k_1 and k_2 are not constant since transcriptional rates are subject to saturation. In models corresponding to Figure 4(b) of the main text and Supplementary Figure 6, k_3 and k_4 are not constant since translational rates are subject to saturation. NA stands for not applicable.

where λ_j ($j = 1-4$) and k_j ($j = 1-4$) are intrinsic rates. Under the hypothesis of transcriptional saturation,

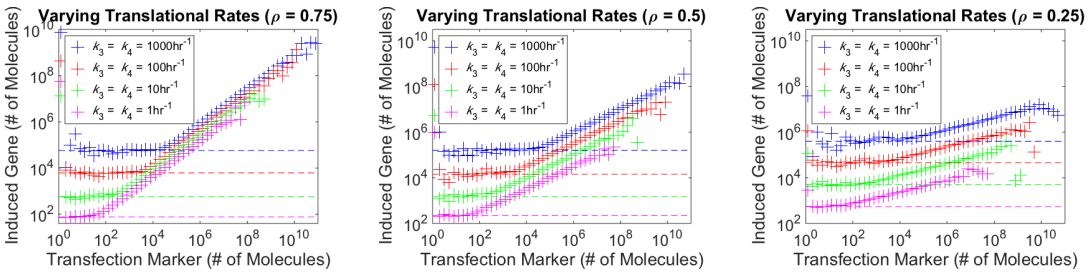
$$K_1 = 1000 \cdot \frac{D_{\text{tm}}}{D_{\text{tm}} + K_{D_{\text{tm}}}},$$
$$K_2 = 1000 \cdot \frac{D_{\text{induced}}}{D_{\text{induced}} + K_{D_{\text{induced}}}},$$

where $K_{D_{\text{tm}}} = 10^4$, and $K_{D_{\text{induced}}} = 10^2, 10^4$, or 10^6. Under the hypothesis of translational saturation,

$$K_3 = 1000000 \cdot \frac{M_{\text{tm}}}{M_{\text{tm}} + 10000},$$
$$K_4 = 1000000 \cdot \frac{M_{\text{induced}}}{M_{\text{induced}} + 10000}.$$

Values of the parameters in each model are shown in Supplementary Table 1.

For the models detailedly described in the main text, the initial total number of plasmids in a given cell is assumed to follow a log-normal distribution: $N[\log(100), \log(10)]$ (Davidsohn et al. (2015)). The initial copy numbers of each species of plasmid, D_{tm} and D_{induced}



Supplementary Figure 3: Comparison of models in which the translational rates decrease in order from 1000 to 1 molecule per mRNA per hour.

given the total number of plasmids P are assumed to follow binomial distributions: $B(P,0.5)$ (Davidsohn et al. (2015)).

2.2 Exploring Other Plasmid Distributions

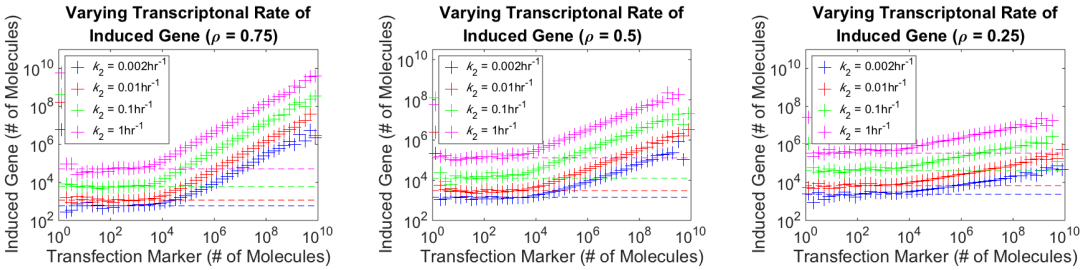
In Davidsohn et al. (2015), co-transfected plasmids were pre-mixed before forming complexes with lipofectamine, and according to Schwake et al. (2010), numbers of co-transfected plasmids in individual cells should be highly correlated. In co-transfection experiments, the correlation between co-transfected plasmids can be adjusted by changing the co-transfection protocol (Schwake et al. (2010)). Besides the models described in the main text, we construct, simulate, and analyze additional cohorts of detailed two-stage models, assuming that numbers of co-transfected plasmids follow a bivariate log-normal distribution, and correlations between co-transfected plasmids can be varied. The initial plasmid copy numbers in a cell, D_{tm} and D_{induced}, are integer roundups of two continuous variables sampled from a bivariate lognormal distribution,

$$N\left[\begin{pmatrix} \log(100) \\ \log(100) \end{pmatrix}, \begin{pmatrix} [\log(10)]^2 & \rho\cdot[\log(10)]^2 \\ \rho\cdot[\log(10)]^2 & [\log(10)]^2 \end{pmatrix}\right]$$

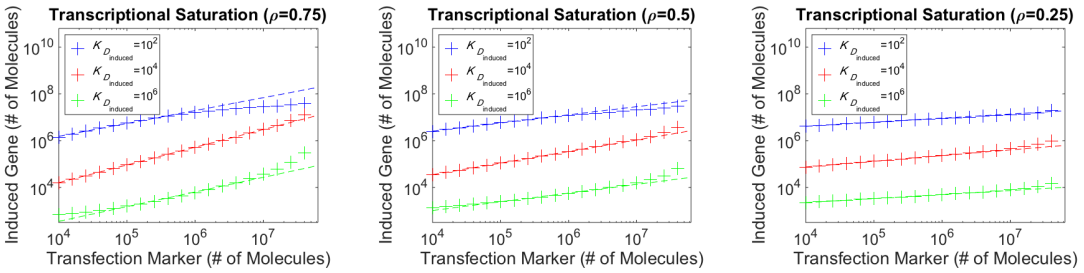
.

ρ represents the correlation between D_{induced} and D_{tm}, and is set to values of 0.25, 0.5, and 0.75 to represent low, medium, and high correlation in different models. Length of the simulation, assumptions about cell division and asynchronicity, and definitions of the reaction rates are kept the same. Values of the rest of the parameters in each model can be found in Supplementary Table 1. Results of the simulation can be found in Supplementary Figures 3, 4, 5, and 6. Irrespective of the underlying plasmid distributions, we reach the same conclusions on biological hypotheses and parameter regions that can explain our experimental observations qualitatively. Another interesting point worth noticing is that as is shown by Supplementary Figures 5 and 6, the saturation behavior is only observed when ρ is set to 0.75, indicating the possible role of co-transfection efficiency as a contributing factor.

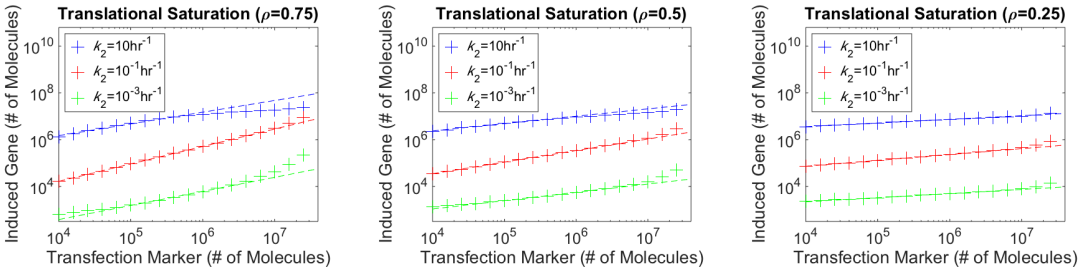
2.3 Reasons for a Near-Constant Level of the Induced Gene Reporter



Supplementary Figure 4: Comparison of models in which the transcriptional rate of the induced gene increases from 0.002 to 1 molecule per plasmid per hour.

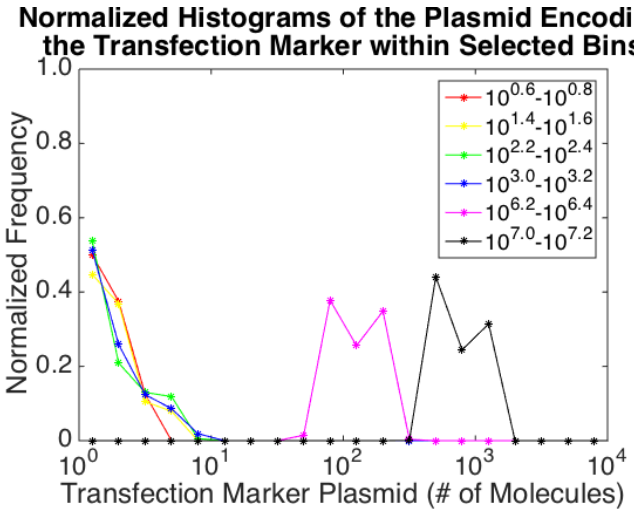


Supplementary Figure 5: Comparison of models built under the hypothesis of transcriptional saturation. The half saturation constant $K_{D_{\text{induced}}}$ increases in order from 10^2 to 10^6 molecules, and $K_{D_{\text{tm}}}$ is held fixed at 10^4 molecules.



Supplementary Figure 6: Comparison of models built under the hypothesis of translational saturation. The transcriptional rate of the induced gene decreases in order from 10 to 10^{-3} molecule per plasmid per hour, and the transfection marker transcribes at a constant rate of 10^{-1} molecule per plasmid per hour.

As is shown in Supplementary Figure 3 (with parameters $k_3 = 1000\text{hr}^{-1}$, $k_4 = 1000\text{hr}^{-1}$, and $\rho = 0.75$), the level of the induced gene reporter stays near a constant value at low plasmid copy numbers. Here no saturation kinetics were included within the model, i.e. the transcription and translation rates were simple linear functions (Supplementary Table 1). We find that at low copy numbers for the plasmid encoding the transfection marker, each of the leftmost bins in Supplementary Figure 7 had relatively constant modes.



Supplementary Figure 7: Normalized histograms of the plasmid encoding the transfection marker within selected bins. The x-axis gives the amount of molecules of the plasmid encoding the transfection marker in units of MEFL, and the y-axis the normalized frequency that amount is observed. Bins between $10^{0.8}$ and $10^{1.0}$ MEFL, $10^{1.4}$ and $10^{1.6}$ MEFL, $10^{2.2}$ and $10^{2.4}$ MEFL, $10^{3.0}$ and $10^{3.2}$ MEFL, $10^{6.2}$ and $10^{6.4}$ MEFL, and $10^{7.0}$ and $10^{7.2}$ MEFL are shown. At low plasmid copy number (low bins), we observe that the distributions of plasmids have almost constant modes; at high copy number (high bins), the modes get shifted to the right as bin index increases.

3 Hill-function-based Models

A Hill function is commonly expressed as:

$$H(I) = \begin{cases} (1-\gamma)\cdot\dfrac{1}{1+\left(\frac{I}{d}\right)^h}+\gamma, & \text{if I is an inhibitor} \\ (1-\gamma)\cdot\dfrac{\left(\frac{I}{d}\right)^h}{1+\left(\frac{I}{d}\right)^h}+\gamma, & \text{if I is an activator,} \end{cases}$$

where I is the concentration of the inhibitor/activator. $H(I)$ accounts for the fraction of the promoter that is active. γ is the minimum fraction of the promoter that is active: if I is an inhibitor, γ is the fraction active given infinite abundance of I; if I is an activator, γ is the fraction active in absence of I. h is the Hill coefficient, and d is the dissociation constant.

The Hill-function-based model we use, originally used in Davidsohn et al. (2015), is:

$$\begin{aligned}
\frac{dI_i}{dt} &= \alpha_i \cdot \phi(t) - \lambda_I \cdot I_i \\
\frac{dO_i}{dt} &= \beta \cdot \phi(t) \cdot \left(\frac{P_i}{P_1}\right)^f \cdot H(I_i) - \lambda_O \cdot O_i \\
\phi(t) &= \left(\frac{1}{2}\right)^{\lfloor \frac{t}{T} \rfloor} \\
H(I_i) &= (1-\gamma) \cdot \frac{1}{1+\left(\frac{I_i}{d}\right)^h} + \gamma,
\end{aligned}$$

where f captures the linear relationship between the maximum production rate of the output protein and the concentration of the transfection marker on the log scale. The rest of the notations follow Equation (1) in the main text.

4 Fitting ODE Models

Assume the regulatory switch is induced at m dosages, and cells are segmented into n bins by their plasmid copy numbers. Let O_{iu} denote the averaged measurements of concentrations of the regulated protein in the i-th bin at dose level u at the final time point t^*, and $\hat{O}_{iu}$ the counterpart numerically simulated by the model. We fit $\log\left(\hat{O}_{iu}\right)$ to $\log(O_{iu})$ by iteratively searching for the set of parameters that minimize the mean-squared error (Carpenter (1960)):

$$\frac{\sum_{u=1}^{m}\sum_{i=1}^{n}\left[\log(O_{iu}) - \log\left(\hat{O}_{iu}\right)\right]^2}{mn - \#\text{ of params}}$$

via the `GlobalSearch` solver in Matlab. `GlobalSearch` uses a scatter-search mechanism to generate start points, initiates a local solver from these start points, and reevaluates the start points during the minimization process. We implement `GlobalSearch` using the local solver `fmincon`, and for `fmincon`, we use the '`sqp`' algorithm. To fit the traditional Hill-function-based models, we set boundaries of $\log_{10}(d)$, $\log_{10}(\beta)$, $\log_{10}(f)$, $\log_{10}(h)$, and $\log_{10}(\gamma)$ to be $[2,8]$, $[-2,6]$, $[-1,1]$, $[-4,4]$, and $[-5,0]$, respectively. To fit the bin-dependent models, we keep the above settings and set the boundary of $\log_{10}(g)$ to be $[-3,1]$. The rest of the search algorithm parameters are set to their default values. The optimized fits are listed in Supporting Information Section 5.

In cross-validation, the fitting errors are defined as (Carpenter (1960)):

$$\frac{\sum_{u=1}^{m}\sum_{i=1}^{n}\left[\log(O_{iu}) - \log\left(\hat{O}_{iu}\right)\right]^2}{mn - \#\text{ of params}}.$$

The testing errors are defined as (Carpenter (1960)):

$$\frac{\sum_{u=1}^{m}\sum_{i=1}^{n}\left[\log(O_{iu}) - \log\left(\hat{O}_{iu}\right)\right]^2}{mn}.$$

In the “Modular Composition" subsection of the main text, the mean-fold error we adopt for evaluating model performance on each cascade and feed-forward circuit is defined as:

$$\frac{\sum_{u=1}^{m}\sum_{i=1}^{n}\left|\log\left(O_{iu}\right)-\log\left(\hat{O}_{iu}\right)\right|}{mn}.$$

5 Optimal Parameter Fits of ODE Models

Optimized fits						
Model	β (Unit: MEFL/hr)	f	d (Unit: MEFL)	h	γ	Error
TAL14	5.52×10^{4}	1.47	1.04×10^{5}	0.73	1.50×10^{-3}	0.013
TAL21	6.96×10^{4}	1.28	2.13×10^{5}	0.68	1.91×10^{-5}	0.015
LmrA	1.51×10^{4}	1.72	2.34×10^{6}	0.92	5.85×10^{-4}	0.020

Supplementary Table 2: Optimal parameters and mean-squared errors for the traditional Hill-function-based model fit to the experimental data. All parameter values are rounded to two digits after the decimal point. The experimental data are from Davidsohn et al. (2015).

Optimized fits					
Model	β (Unit: MEFL/hr)	f	d (Unit: MEFL)	h	Error
TAL14	4.87×10^{4}	1.74	5.39×10^{4}	0.68	0.004
TAL21	4.68×10^{4}	1.58	2.90×10^{5}	0.72	0.005
LmrA	1.66×10^{4}	1.91	3.73×10^{5}	0.59	0.009
	γ	g			
TAL14	2.83×10^{-4}	1.10			
TAL21	1.10×10^{-3}	0.83			
LmrA	2.36×10^{-5}	1.09			

Supplementary Table 3: Optimal parameters and mean-squared errors for the bin-dependent model fit to the experimental data. All parameter values are rounded to two digits after the decimal point. The experimental data are from Davidsohn et al. (2015).

6 Models for Cascades

Upon modular connection, parameters that are fit to input-output curves of individual modules need to be corrected for batch effects. As is shown in Supporting Information Section 12 of Davidsohn et al. (2015), the rescaling factors for the input protein I, the output protein O, and the transfection marker are TAL14: 0.29, 0.93, 0.89; TAL21: 0.20, 1, 1.12; LmrA: 1, 0.41, 1 (Davidsohn et al. (2015)). For example, for output protein O, TAL14 has a scaling factor of 0.93, and TAL21, a factor of 1. This means to compare the output protein between TAL14 and TAL21, data for TAL14 need to be multiplied by 0.93 so that the two are brought to the same scale. The scaling factors are used to rescale the parameters in the bin-dependent models before the models are connected into a chain. d is rescaled with the input, β rescaled with the output, and P_i rescaled with the transfection marker. Mathematically speaking, if c_I, c_O, and c_P are the scaling factors of the input, the output, and the transfection marker, then the rescaled bin-dependent model is formulated as follows:

$$
\begin{aligned}
\frac{dI'_i}{dt} &= \alpha'_i \cdot \phi(t) - \lambda \cdot I'_i, \\
\frac{dO'_i}{dt} &= \begin{cases}
\beta' \cdot \phi(t) \cdot \left(\frac{P'_i}{P'_1}\right)^f \cdot \left(\frac{1-\gamma}{1+\left(\frac{I'_i}{d'}\right)^h} + \gamma\right) - \lambda \cdot O'_i, & \text{if } P'_i < P'_{i'} \\
\beta' \cdot \phi(t) \cdot \left(\frac{P'_{i'}}{P'_1}\right)^f \cdot \left(\frac{P'_i}{P'_{i'}}\right)^g \cdot \frac{1-\gamma}{1+\left(\frac{I'_i}{d'}\right)^h} & \\
\quad + \beta' \cdot \phi(t) \cdot \left(\frac{P'_i}{P'_1}\right)^f \cdot \gamma - \lambda \cdot O'_i, & \text{if } P'_i \geq P'_{i'}
\end{cases} \\
\phi(t) &= \left(\frac{1}{2}\right)^{\lfloor \frac{t}{T} \rfloor},
\end{aligned}
$$

where the prime variables represent the variables without batch effects:

$$
\begin{aligned}
I'_i &= I_i \cdot c_I & \beta' &= \beta \cdot c_O \cdot c_P & P'_i &= P_i \cdot c_P & \alpha'_i &= \alpha_i \cdot c_I \\
O'_i &= O_i \cdot c_O \cdot c_P & d' &= d \cdot c_I & P'_{i'} &= P_{i'} \cdot c_P. &&
\end{aligned}
$$

As is shown in Supporting Information Section 12 of Davidsohn et al. (2015), scaling factors of the transfection marker for the cascades, $\tilde{c}_P$ are $\{1.51, 1.07, 0.68, 0.78, 0.71, 0.79\}$ for TAL14-TAL21, TAL14-LmrA, TAL21-TAL14, TAL21-LmrA, LmrA-TAL14, and LmrA-TAL21, respectively. For all these cascades, $\tilde{c}_I = 1$, and $\tilde{c}_O = 1$. Since the prime variables involve no batch effects, to convert to a cascade, we must divide all the prime variables by the corresponding cascade scaling factors ($\tilde{c}_I$, $\tilde{c}_O$, $\tilde{c}_P$). In addition, to offer a comparable study to Davidsohn et al. (2015), we follow similar implementation details as are shown in Davidsohn et al. (2015) by multiplying the dissociation constant of the the second repressor by three (see the fourth to last paragraph of the Supporting Information Section 5 of Davidsohn et al. (2015)). This is because the plasmids for the second repressor are transfected at one-third the concentration of

the first repressor (Davidsohn et al. (2015)). This suggests that production of the second repressor should scale like one-third the activation level of the first repressor during the initial transient, when much of the repressor is produced for the system (Davidsohn et al. (2015)). The final bin-dependent model for cascades is expressed as:

$$\frac{dI_i''}{dt} = \alpha_i'' \cdot \phi(t) - \lambda \cdot I_i''$$

$$\frac{dO_{1i}''}{dt} = \begin{cases} \beta_1'' \cdot \phi(t) \cdot \left(\frac{P_{1i}''}{P_{11}''}\right)^{f_1} \cdot \left(\frac{1-\gamma_1}{1+\left(\frac{I_i''}{d_1''}\right)^{h_1}} + \gamma_1\right) - \lambda \cdot O_{1i}'', & \text{if } P_{1i}'' < P_{1i'}'' \\ \beta_1'' \cdot \phi(t) \cdot \left(\frac{P_{1i'}''}{P_{11}''}\right)^{f_1} \cdot \left(\frac{P_{1i}''}{P_{1i'}''}\right)^{g_1} \cdot \frac{1-\gamma_1}{1+\left(\frac{I_i''}{d_1''}\right)^{h_1}} & \\ \quad + \beta_1'' \cdot \phi(t) \cdot \left(\frac{P_{1i}''}{P_{11}''}\right)^{f_1} \cdot \gamma_1 - \lambda \cdot O_{1i}'', & \text{if } P_{1i}'' \geq P_{1i'}'' \end{cases}$$

$$\frac{dO_{2i}''}{dt} = \begin{cases} \beta_2'' \cdot \phi(t) \cdot \left(\frac{P_{2i}''}{P_{21}''}\right)^{f_2} \cdot \left(\frac{1-\gamma_2}{1+\left(\frac{O_{1i}''}{3 \cdot d_2''}\right)^{h_2}} + \gamma_2\right) - \lambda \cdot O_{2i}'', & \text{if } P_{2i}'' < P_{2i'}'' \\ \beta_2'' \cdot \phi(t) \cdot \left(\frac{P_{2i'}''}{P_{21}''}\right)^{f_2} \cdot \left(\frac{P_{2i}''}{P_{2i'}''}\right)^{g_2} \cdot \frac{1-\gamma_2}{1+\left(\frac{O_{1i}''}{3 \cdot d_2''}\right)^{h_2}} & \\ \quad + \beta_2'' \cdot \phi(t) \cdot \left(\frac{P_{2i}''}{P_{21}''}\right)^{f_2} \cdot \gamma_2 - \lambda \cdot O_{2i}'', & \text{if } P_{2i}'' \geq P_{2i'}'' \end{cases}$$

$$\phi(t) = \left(\frac{1}{2}\right)^{\lfloor \frac{t}{T} \rfloor},$$

where

$$I_i'' = \frac{I_i'}{\tilde{c}_I} \quad \alpha_i'' = \frac{\alpha_i'}{\tilde{c}_I},$$

and for the k-th module ($k = 1, 2$) and the j-th cascade ($j = 1 - 6$),

$$\beta_k'' = \frac{\beta_k'}{\tilde{c}_{Pj} \cdot \tilde{c}_O} \quad P_{ki}'' = \frac{P_{ki}'}{\tilde{c}_{Pj}} \quad d_k'' = \frac{d_k'}{\tilde{c}_I}$$
$$O_{ki}'' = \frac{O_{ki}'}{\tilde{c}_{Pj} \cdot \tilde{c}_O} \quad P_{ki'}'' = \frac{P_{ki'}'}{\tilde{c}_{Pj}}.$$

The double prime variables represent variables that account for the batch effects of the cascades. Values of the parameters used in the final bin-dependent models for six cascades are shown in Supplementary Table 4.

The same method of rescaling can be applied to the Hill-function-based model presented in Davidsohn et al. (2015). The rescaled Hill-function model for cascades is expressed as:

$$\begin{aligned}
\frac{dI_i''}{dt} &= \alpha_i'' \cdot \phi(t) - \lambda \cdot I_i'' \\
\frac{dO_{1i}''}{dt} &= \beta_1'' \cdot \phi(t) \cdot \left(\frac{P_{1i}''}{P_{11}''}\right)^{f_1} \cdot \left(\frac{1-\gamma_1}{1+\left(\frac{I_i''}{d_1''}\right)^{h_1}} + \gamma_1\right) - \lambda \cdot O_{1i}'', \\
\frac{dO_{2i}''}{dt} &= \beta_2'' \cdot \phi(t) \cdot \left(\frac{P_{2i}''}{P_{21}''}\right)^{f_2} \cdot \left(\frac{1-\gamma_2}{1+\left(\frac{O_{1i}''}{3\cdot d_2''}\right)^{h_2}} + \gamma_2\right) - \lambda \cdot O_{2i}'', \\
\phi(t) &= \left(\frac{1}{2}\right)^{\lfloor \frac{t}{T} \rfloor},
\end{aligned}$$

where

$$I_i'' = \frac{I_i'}{\tilde{c}_I} \quad \alpha_i'' = \frac{\alpha_i'}{\tilde{c}_I},$$

and for the k-th module ($k = 1, 2$) and the j-th cascade ($j = 1 - 6$),

$$\begin{aligned}
\beta_k'' &= \frac{\beta_k'}{\tilde{c}_{Pj} \cdot \tilde{c}_O} \quad & P_{ki}'' &= \frac{P_{ki}'}{\tilde{c}_{Pj}} \\
d_k'' &= \frac{d_k'}{\tilde{c}_I} & O_{ki}'' &= \frac{O_{ki}'}{\tilde{c}_{Pj} \cdot \tilde{c}_O}.
\end{aligned}$$

Values of the parameters used in the rescaled Hill-function-based models for six cascades are shown in Supplementary Table 5.

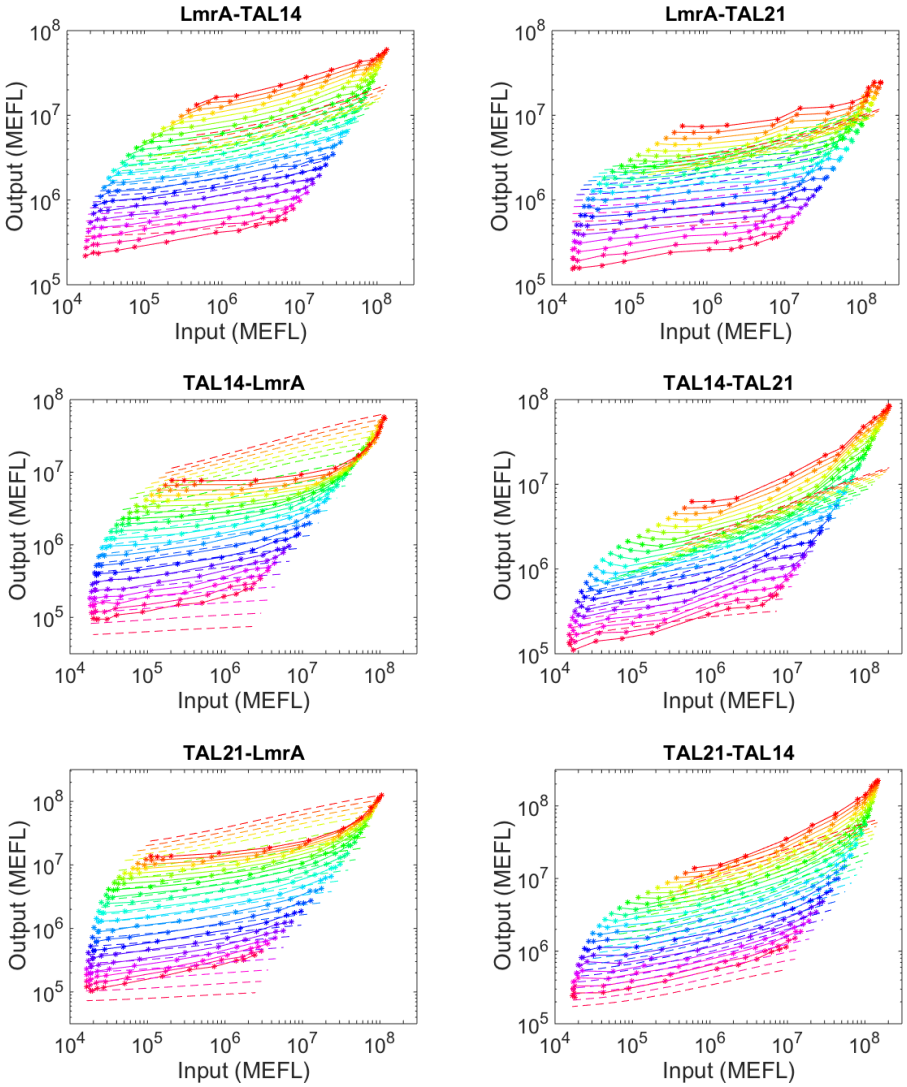
Parameters					
Cascade	β_1'' (MEFL/hr)	f_1	d_1'' (MEFL)	h_1	γ_1
LmrA-TAL14	9.78×10^3	1.91	3.73×10^5	0.59	2.36×10^{-5}
LmrA-TAL21	8.67×10^3	1.91	3.73×10^5	0.59	2.36×10^{-5}
TAL14-LmrA	4.70×10^4	1.74	1.58×10^4	0.68	2.83×10^{-4}
TAL14-TAL21	3.34×10^4	1.74	1.58×10^4	0.68	2.83×10^{-4}
TAL21-LmrA	5.34×10^4	1.58	5.77×10^4	0.72	1.10×10^{-3}
TAL21-TAL14	6.20×10^4	1.58	5.77×10^4	0.72	1.10×10^{-3}
	g_1	β_2'' (MEFL/hr)	f_2	d_2'' (MEFL)	h_2
LmrA-TAL14	1.09	7.24×10^4	1.74	1.58×10^4	0.68
LmrA-TAL21	1.09	5.28×10^4	1.58	5.77×10^4	0.72
TAL14-LmrA	1.10	6.35×10^3	1.91	3.73×10^5	0.59
TAL14-TAL21	1.10	2.75×10^4	1.58	5.77×10^4	0.72
TAL21-LmrA	0.83	8.76×10^3	1.91	3.73×10^5	0.59
TAL21-TAL14	0.83	7.54×10^4	1.74	1.58×10^4	0.68
	γ_2	g_2	λ (hr^{-1})	$P_{1i'}''$ (MEFL)	$P_{2i'}''$ (MEFL)
LmrA-TAL14	2.83×10^{-4}	1.10	3.41×10^{-2}	$10^{7.55}$	$10^{7.31}$
LmrA-TAL21	1.10×10^{-3}	0.83	3.41×10^{-2}	$10^{7.51}$	$10^{7.15}$
TAL14-LmrA	2.36×10^{-5}	1.09	3.41×10^{-2}	$10^{7.12}$	$10^{7.37}$
TAL14-TAL21	1.10×10^{-3}	0.83	3.41×10^{-2}	$10^{6.97}$	$10^{6.87}$
TAL21-LmrA	2.36×10^{-5}	1.09	3.41×10^{-2}	$10^{7.16}$	$10^{7.51}$
TAL21-TAL14	2.83×10^{-4}	1.10	3.41×10^{-2}	$10^{7.22}$	$10^{7.32}$

Supplementary Table 4: Values of the rescaled parameters used in the final bin-dependent models for the six cascades.

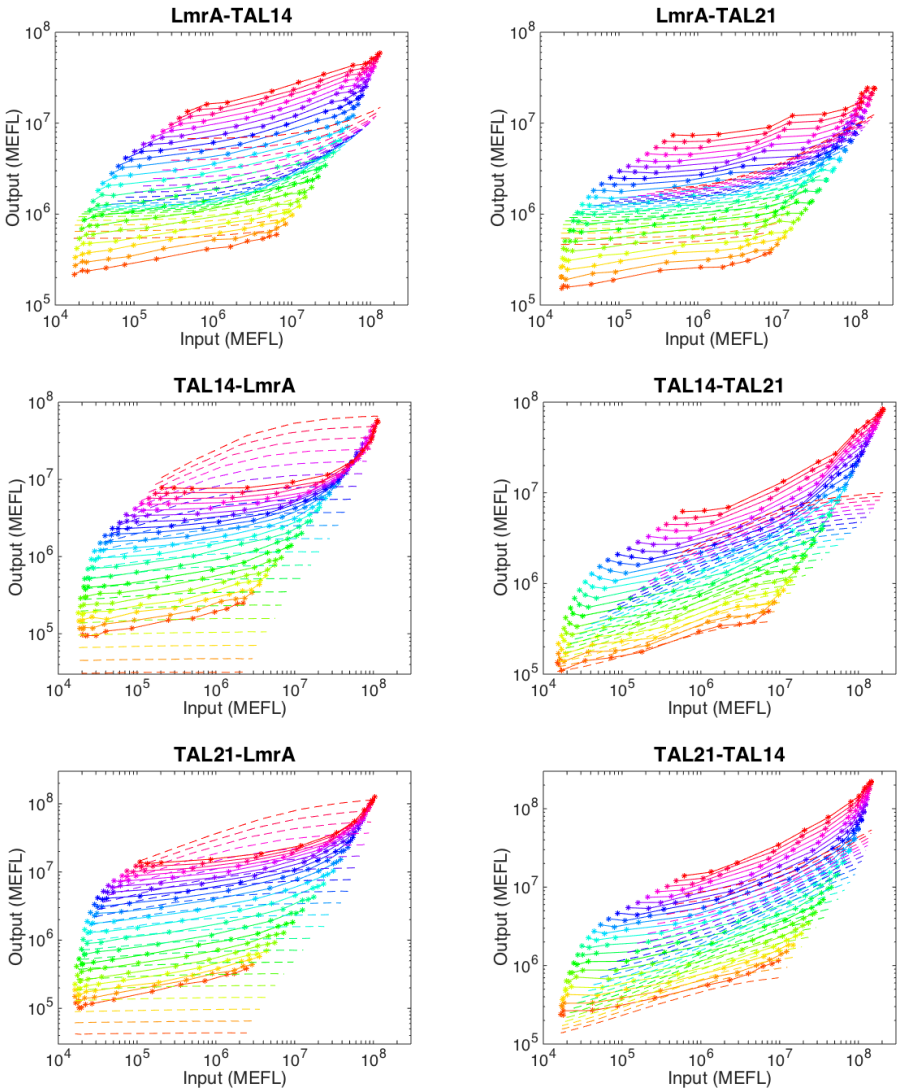
Parameters					
Cascade	β_1'' (MEFL/hr)	f_1	d_1'' (MEFL)	h_1	γ_1
LmrA-TAL14	5.20×10^3	1.91	3.39×10^6	1.00	5.90×10^{-3}
LmrA-TAL21	4.61×10^3	1.91	3.39×10^6	1.00	5.90×10^{-3}
TAL14-LmrA	5.64×10^4	1.45	3.29×10^4	0.98	9.30×10^{-3}
TAL14-TAL21	4.01×10^4	1.45	3.29×10^4	0.98	9.30×10^{-3}
TAL21-LmrA	1.10×10^5	1.74	1.09×10^4	0.66	9.33×10^{-8}
TAL21-TAL14	1.28×10^5	1.74	1.09×10^4	0.66	9.33×10^{-8}
	β_2'' (MEFL/hr)	f_2	d_2'' (MEFL)	h_2	γ_2
LmrA-TAL14	8.69×10^4	1.45	3.29×10^4	0.98	9.30×10^{-3}
LmrA-TAL21	1.09×10^5	1.74	1.09×10^4	0.66	9.33×10^{-8}
TAL14-LmrA	3.38×10^3	1.91	3.39×10^6	1.00	5.90×10^{-3}
TAL14-TAL21	5.69×10^4	1.74	1.09×10^4	0.66	9.33×10^{-8}
TAL21-LmrA	4.66×10^3	1.91	3.39×10^6	1.00	5.90×10^{-3}
TAL21-TAL14	9.04×10^4	1.45	3.29×10^4	0.98	9.30×10^{-3}
	λ (hr^{-1})				
LmrA-TAL14	3.41×10^{-2}				
LmrA-TAL21	3.41×10^{-2}				
TAL14-LmrA	3.41×10^{-2}				
TAL14-TAL21	3.41×10^{-2}				
TAL21-LmrA	3.41×10^{-2}				
TAL21-TAL14	3.41×10^{-2}				

Supplementary Table 5: Values of the rescaled parameters used in the rescaled Hill-function models for the six cascades.

7 Cascade Predictions



Supplementary Figure 8: Comparison between experimental data and predictions made by the bin-dependent model for six cascades. Plasmid copy number is shown by color. Solid lines are experimental data, and dashed lines are model predictions. The experimental data are from Davidsohn et al. (2015).



Supplementary Figure 9: Comparison between experimental data and predictions made by the rescaled Hill-function-based model for six cascades. Plasmid copy number is shown by color. Solid lines are experimental data, and dashed lines are model predictions. The experimental data are from Davidsohn et al. (2015).

8 Models for Feed-forward Circuits

Similar to cascades, the bin-dependent model for feed-forward circuits can be constructed and is expressed as:

$$\frac{dI'_i}{dt} = \alpha'_i \cdot \phi(t) - \lambda \cdot I'_i$$

$$\frac{dO'_{1i}}{dt} = \begin{cases} \beta'_1 \cdot \phi(t) \cdot \left(\frac{P'_{1i}}{P'_{11}}\right)^{f_1} \cdot \left(\frac{1-\gamma_1}{1+\left(\frac{I'_i}{d'_1}\right)^{h_1}} + \gamma_1\right) - \lambda \cdot O'_{1i}, & \text{if } P'_{1i} < P'_{1i'} \\ \beta'_1 \cdot \phi(t) \cdot \left(\frac{P'_{1i'}}{P'_{11}}\right)^{f_1} \cdot \left(\frac{P'_{1i}}{P'_{1i'}}\right)^{g_1} \cdot \frac{1-\gamma_1}{1+\left(\frac{I'_i}{d'_1}\right)^{h_1}} & \\ \quad + \beta'_1 \cdot \phi(t) \cdot \left(\frac{P'_{1i}}{P'_{11}}\right)^{f_1} \cdot \gamma_1 - \lambda \cdot O'_{1i}, & \text{if } P'_{1i} \geq P'_{1i'} \end{cases}$$

$$\frac{dO'_{2i}}{dt} = \begin{cases} \beta'_2 \cdot \phi(t) \cdot \left(\frac{P'_{2i}}{P'_{21}}\right)^{f_2} \cdot \left(\frac{1-\gamma_2}{1+\left(\frac{3 \cdot I'_i + O'_{1i}}{3 \cdot d'_2}\right)^{h_2}} + \gamma_2\right) - \lambda \cdot O'_{2i}, & \text{if } P'_{2i} < P'_{2i'} \\ \beta'_2 \cdot \phi(t) \cdot \left(\frac{P'_{2i'}}{P'_{21}}\right)^{f_2} \cdot \left(\frac{P'_{2i}}{P'_{2i'}}\right)^{g_2} \cdot \frac{1-\gamma_2}{1+\left(\frac{3 \cdot I'_i + O'_{1i}}{3 \cdot d'_2}\right)^{h_2}} & \\ \quad + \beta'_2 \cdot \phi(t) \cdot \left(\frac{P'_{2i}}{P'_{21}}\right)^{f_2} \cdot \gamma_2 - \lambda \cdot O'_{2i}, & \text{if } P'_{2i} \geq P'_{2i'} \end{cases}$$

$$\phi(t) = \left(\frac{1}{2}\right)^{\lfloor \frac{t}{T} \rfloor},$$

where the prime variables represent the variables without batch effects. The first equation involving I'_i describes the dynamics of I_1 as well as I_3, which directly inhibits O_2 (Figure 10(a)). Note that due to uncontrollable experimental variations, the distributions of the transfection marker differ substantially between feed-forward circuits and rest of the circuits studied in this paper (see Supporting Information Section 6 of Davidsohn et al. (2015)). The fluorescence-intensity distributions of the transfection marker reach the upper ends near $10^{7.4}$ MEFL for LmrA-TAL14 and TAL21-LmrA feed-forward circuits, and near $10^{8.0}$ MEFL for TAL21-TAL14 feed-forward circuit, while the upper ends lie near $10^{8.5}$ for modules as well as cascades. To reconcile the differences, we divide $P'_{1i'}$ and $P'_{2i'}$ in the feed-forward circuit models by the fold differences between the upper ends of fluorescence-intensity distributions of the transfection marker for modules and feed-forward circuits. Values of parameters used in the actual models, including $P'_{1i'}$ and $P'_{2i'}$, can be found in Supplementary Table 6.

The rescaled Hill-function model for feed-forward circuits is expressed as:

$$\frac{dI'_i}{dt} = \alpha'_i \cdot \phi(t) - \lambda \cdot I'_i$$

$$\frac{dO'_{1i}}{dt} = \beta'_1 \cdot \phi(t) \cdot \left(\frac{P'_{1i}}{P'_{11}}\right)^{f_1} \cdot \left(\frac{1-\gamma_1}{1+\left(\frac{I'_i}{d'_1}\right)^{h_1}} + \gamma_1\right) - \lambda \cdot O'_{1i},$$

$$\frac{dO'_{2i}}{dt} = \beta'_2 \cdot \phi(t) \cdot \left(\frac{P'_{2i}}{P'_{21}}\right)^{f_2} \cdot \left(\frac{1-\gamma_2}{1+\left(\frac{3\cdot I'_i + O'_{1i}}{3\cdot d'_2}\right)^{h_2}} + \gamma_2\right) - \lambda \cdot O'_{2i},$$

$$\phi(t) = \left(\frac{1}{2}\right)^{\lfloor \frac{t}{T} \rfloor}$$

.

Values of parameters used in the actual Hill-function models can be found in Supplementary Table 7.

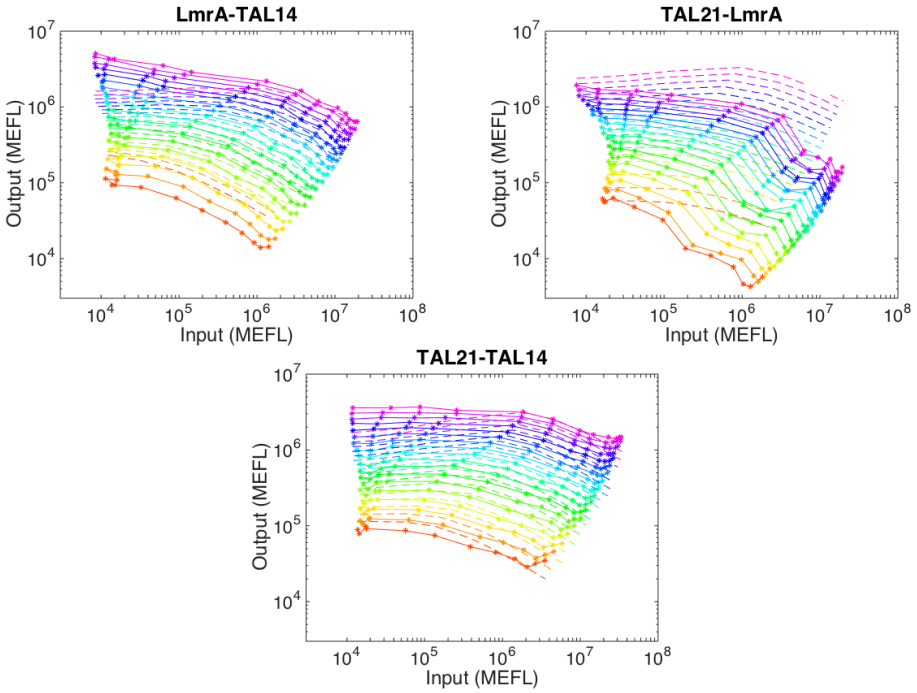
Parameters					
Circuit	β_1' (MEFL/hr)	f_1	d_1' (MEFL)	h_1	γ_1
LmrA-TAL14	6.86×10^3	1.91	3.73×10^5	0.59	2.36×10^{-5}
TAL21-LmrA	4.68×10^4	1.58	5.77×10^4	0.72	1.10×10^{-3}
TAL21-TAL14	4.68×10^4	1.58	5.77×10^4	0.72	1.10×10^{-3}
	g_1	β_2' (MEFL/hr)	f_2	d_2' (MEFL)	h_2
LmrA-TAL14	1.09	4.53×10^4	1.74	1.58×10^4	0.68
TAL21-LmrA	0.83	6.86×10^3	1.91	3.73×10^5	0.59
TAL21-TAL14	0.83	4.53×10^4	1.74	1.58×10^4	0.68
	γ_2	g_2	λ (hr^{-1})	$P_{1i'}'$ (MEFL)	$P_{2i'}'$ (MEFL)
LmrA-TAL14	2.83×10^{-4}	1.10	3.41×10^{-2}	$10^{6.25}$	$10^{5.95}$
TAL21-LmrA	2.36×10^{-5}	1.09	3.41×10^{-2}	$10^{5.95}$	$10^{6.25}$
TAL21-TAL14	2.83×10^{-4}	1.10	3.41×10^{-2}	$10^{6.55}$	$10^{6.55}$

Supplementary Table 6: Values of the rescaled parameters used in the bin-dependent models for the three feed-forward circuits.

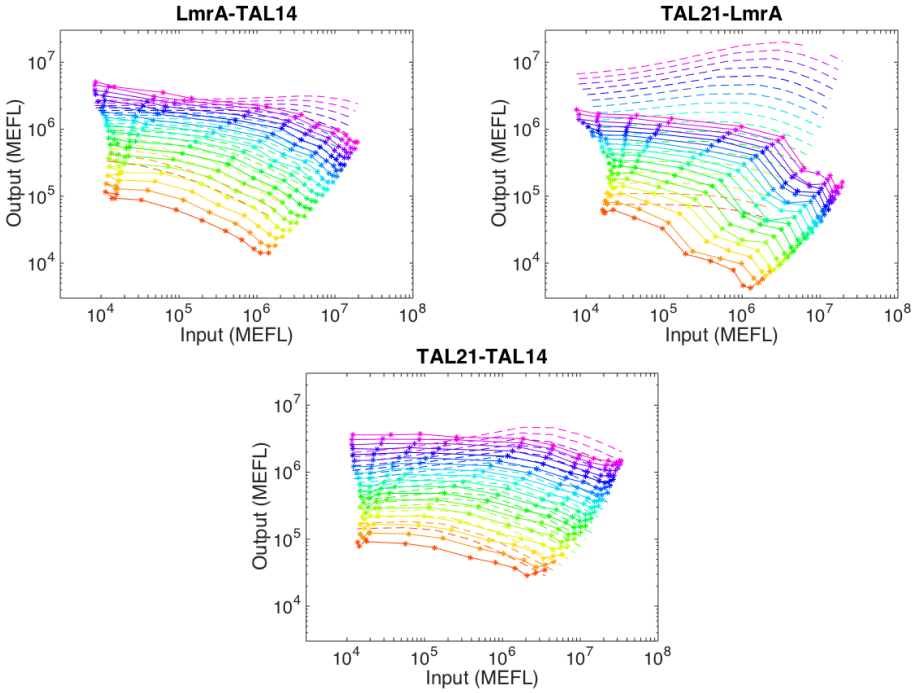
Parameters					
Circuit	β_1' (MEFL/hr)	f_1	d_1' (MEFL)	h_1	γ_1
LmrA-TAL14	6.24×10^3	1.72	2.34×10^6	0.92	5.85×10^{-4}
TAL21-LmrA	6.96×10^4	1.28	4.25×10^4	0.68	1.91×10^{-5}
TAL21-TAL14	6.96×10^4	1.28	4.25×10^4	0.68	1.91×10^{-5}
	β_2' (MEFL/hr)	f_2	d_2' (MEFL)	h_2	γ_2
LmrA-TAL14	5.14×10^4	1.47	3.04×10^4	0.73	1.50×10^{-3}
TAL21-LmrA	6.24×10^3	1.72	2.34×10^6	0.92	5.85×10^{-4}
TAL21-TAL14	5.14×10^4	1.47	3.04×10^4	0.73	1.50×10^{-3}
	λ (hr^{-1})				
LmrA-TAL14	3.41×10^{-2}				
TAL21-LmrA	3.41×10^{-2}				
TAL21-TAL14	3.41×10^{-2}				

Supplementary Table 7: Values of the rescaled parameters used in the rescaled Hill-function models for the three feed-forward circuits.

9 Feed-forward Circuit Predictions



Supplementary Figure 10: Comparison between experimental data and predictions made by the bin-dependent model for three feed-forward circuits. Plasmid copy number is shown by color. Solid lines are experimental data, and dashed lines are model predictions. The experimental data are from Davidsohn et al. (2015).



Supplementary Figure 11: Comparison between experimental data and predictions made by the rescaled Hill-function-based model for three feed-forward circuits. Plasmid copy number is shown by color. Solid lines are experimental data, and dashed lines are model predictions. The experimental data are from Davidsohn et al. (2015).

References

Beal, J. (2017). Biochemical complexity drives log-normal variation in genetic expression. *Engineering Biology*, 1(1):55–60.

Brahme, A., editor (2014). *Comprehensive Biomedical Physics*. Elsevier.

Carpenter, R. G. (1960). *Principles and Procedures of Statistics, with Special Reference to the Biological Sciences*, volume 52. McGraw-Hill.

Davidsohn, N., Beal, J., Kiani, S., Adler, A., Yaman, F., Li, Y., Xie, Z., and Weiss, R. (2015). Accurate predictions of genetic circuit behavior from part characterization and modular composition. *ACS Synth. Biol.*, 4(6):673–681.

Schwake, G., Youssef, S., Kuhr, J.-T., Gude, S., David, M. P., Mendoza, E., Frey, E., and Radler, J. O. (2010). Predictive modeling of non-viral gene transfer. *Biotechnol. Bioeng.*, 105(4):805–813.