Population Design for Synthetic Gene Circuits

Baptiste Turpin, Eline Y. Bijman, Hans-Michael Kaltenbach,
and Jörg Stelling[(✉)]

Department of Biosystems Science and Engineering (D-BSSE) and SIB Swiss
Institute of Bioinformatics, ETH Zurich, 4058 Basel, Switzerland
`joerg.stelling@bsse.ethz.ch`

Abstract. Synthetic biologists use and combine diverse biological parts
to build systems such as genetic circuits that perform desirable functions
in, for example, biomedical or industrial applications. Computer-aided
design methods have been developed to help choose appropriate network
structures and biological parts for a given design objective. However,
they almost always model the behavior of the network in an average
cell, despite pervasive cell-to-cell variability. Here, we present a compu-
tational framework to guide the design of synthetic biological circuits
while accounting for cell-to-cell variability explicitly. Our design method
integrates a NonLinear Mixed-Effect (NLME) framework into an exist-
ing algorithm for design based on ordinary differential equation (ODE)
models. The analysis of a recently developed transcriptional controller
demonstrates first insights into design guidelines when trying to achieve
reliable performance under cell-to-cell variability. We anticipate that our
method not only facilitates the rational design of synthetic networks
under cell-to-cell variability, but also enables novel applications by sup-
porting design objectives that specify the desired behavior of cell popu-
lations.

Keywords: Cell-to-cell variability · Synthetic biology ·
Computer-aided design

1 Introduction

Synthetic biology aims at establishing novel functions in biological systems, or
to re-engineer existing ones, in many areas such as new materials or cell-based
therapies that are starting to see real-world applications [21]. The conceptual
core of the field's rational engineering approach to establish, for example, the
corresponding synthetic gene circuits are a systematic design-build-test cycle
and the use of predictive mathematical models throughout this cycle to design,
analyze, and tune the circuits [14].

Computer-aided design helps identifying suitable network structures (topolo-
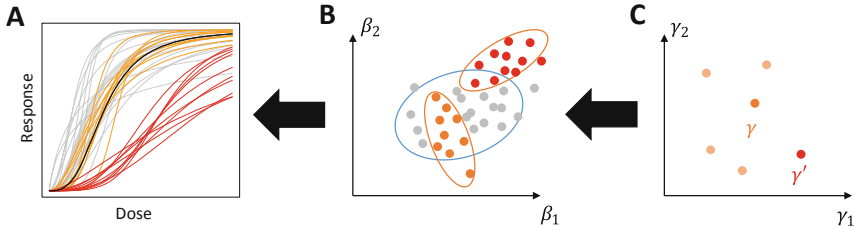gies) as well as biological parts for their implementation to reach a given design

Fig. 1. Cell behaviors relate to parameters at the individual and population level. (A) Dose-response relationships for single cells (lines) drawn from two distinct populations (red and orange) as well as other cells (gray). The design objective for individual cells is represented by an ideal reference curve (black). **(B)** Space of individual parameters β, the set of possible parameter values for a single cell. Dots show parametrizations yielding the behaviors in **(A)** of the corresponding color. The blue ellipse encloses the individual viable space where an individual cost measuring consistency of the single-cell behavior with the design objective for individual cells is below a threshold ε. Red and orange dots encircled by ellipses represent individual cells drawn from the two distinct cell populations. **(C)** Space of population parameters γ, where each parameter vector (dot) describes a full distribution of individual parameters in a population, typically via mean vector and covariance matrix. The orange (γ) and red (γ') dots represent the population parameter vectors that generate the corresponding populations in **(A,B)**. (Color figure online)

objective. For the commonly applied models in the form of ordinary differential equations (ODEs), both design problems can be addressed by investigating the space of model parameters to assess (predicted) circuit behaviors in relation to design objectives encoded by a reference for the desired behavior. With sampling-based methods such as (approximate) Bayesian computation, this defines a 'viable' subspace of the parameter space where the behavior is consistent with the design objective (Fig. 1**A,B**) [2,10,17].

The ODE-based approach captures the behavior of an 'average' cell and thus only allows design with respect to such an assumed cell. Yet, for the biological implementation it is critical that a circuit functions under conditions of uncertainty (e.g., in changing environmental conditions or because the models do not capture relevant interactions between parts or with the cellular context [7]) as well as cell-to-cell variability that is present even in isogenic populations (e.g., due to extrinsic or intrinsic stochastic noise, or different cell cycle phases and ages of cells in a population [4]). One can account for uncertainty in ODE-based design, for example, via measures of robustness that quantify parameter uncertainty [10]. It is also possible to tackle cell-to-cell variability with stochastic models, where temporal logic specifications are written as Continuous Stochastic Logic (CSL) [23]. However, the pure ODE and CSL frameworks are limited in two main aspects: First, they cannot account for all aspects of cell-to-cell variability directly; stochastic models do not represent extrinsic variability resulting, for example, from variable cell sizes. This is particularly important when

an 'average' cell poorly represents the population dynamics, for example, when subpopulations of cells show different qualitative behaviors. Second, and related, it is not possible to define design objectives for the population, such as requiring a certain fraction of the cells to have a coherent behavior.

To address these limitations, here we propose a framework for robust synthetic circuit design that takes into account cell-to-cell variability, and clearly separates it from experimental noise and impact of variable environmental conditions and interacting parts. For this population design, we extend an existing algorithm for ODE-based design [10] to the NLME (NonLinear Mixed-Effect) models framework [8]. Specifically, this entails augmenting the ODE model with a statistical model at the population level that induces probability distributions over the parameter space at the individual cell level (see Fig. 1**B,C**). This allows a designer to impose cell-to-cell variability constraints on synthetic networks. We demonstrate the approach with the a posteriori analysis of a recently developed transcriptional controller [1], a class of circuits that is often designed to minimize cell-to-cell variability.

2   Population Design Framework

Individual Cell Model. For any individual cell, the dynamics of the synthetic circuit are governed by the *individual cell model*

$$\Sigma(\beta) : \begin{cases} \frac{dx(t)}{dt} = v(x(t), u(t), \alpha) \\ x(0) = x^0 \\ y(t) = h(x(t)) \, , \end{cases} \tag{1}$$

where x are the system states such as concentrations of chemical species, v is a rate function, and u is an input function. Usually, states cannot be observed directly and the observations y of the system result from a (known) observation function h. We subsume the parameters α and initial conditions x^0 into the parameter $\beta = (\alpha, x^0) \in B$, where B is a bounded set.

Average Cell Design. We first consider the *average cell design problem* of determining the parameter β^* that minimizes the divergence between the circuit's behavior and a desired reference behavior. We model the *behavior of Σ* as an input-output map $D : \mathbb{R} \times B \times \mathcal{U} \to \mathbb{R}$ that provides a (time-dependent) function $D(\tau; \beta, u)$ in τ for each parameter $\beta \in B$ and any input $u \in \mathcal{U}$, where $\mathcal{U}$ is a finite set of relevant inputs. The *reference behavior* $D^{\mathrm{ref}} : \mathbb{R} \times \mathcal{U} \to \mathbb{R}$ is a user-specified (time-dependent) function for each $u \in \mathcal{U}$ that encodes the desired input-output relation; it need not be realizable by Σ. A simple example is a dose-response curve, where a constant input u is mapped to a constant response for the reference, and to the output at steady state for $t \to \infty$ for the circuit. Another example identifies $D(\tau; \beta, u) = y(\tau)$ as the observations of Σ at time τ for a given input and parameter.

We measure the divergence between system and reference behavior by the *individual cost function*

$$s(\beta) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \left|\left| D(\tau; \beta, u) - D^{\mathrm{ref}}(\tau; u) \right|\right| , \qquad (2)$$

which averages some norm $||\cdot||$ between the system and reference behavior over the considered inputs.

In principle, the average cell design problem could be solved directly to identify the optimal average cell parameter $\beta^* = \mathrm{argmin}_\beta s(\beta)$. However, additional uncertainties arise due to unmodelled system components and from combining previously characterized biological parts into a circuit [11]. We account for these uncertainties by defining a threshold $\varepsilon > 0$ on the cost function to encode which solutions are 'good enough', and determine the *viable region* $V^{\mathrm{avg}} = \{\beta \in B \mid s(\beta) \leq \varepsilon\}$ of all parameters that fulfill this criterion. An output of the average cell design problem is then a description of V^{avg} rather than a single parameter.

Population Model. To capture cell-to-cell variability, we postulate a *population model*, where all cells share the same model structure Σ, but each cell i has its own parameter β_i drawn from a common *population distribution*

$$\beta_i \sim P_\gamma \qquad (3)$$

with *population parameters* $\gamma \in \Gamma$. This is known as a *nonlinear mixed-effects model* and P_γ is often chosen to be a normal or log-normal distribution, in which case γ are the expected values and (co)variances of the parameters in β_i.

Population Design. The population model allows us to consider the distribution of behaviors of a circuit under cell-to-cell heterogeneity. In particular, each population parameter γ yields a specific distribution P_γ of the individual cell parameters β, and this induces a distribution over the values of the individual cost functions $s(\beta)$. The *population design problem* then consists of finding a population parameter that minimizes a corresponding *population cost function*, given by a functional

$$c : \{P_\gamma \mid \gamma \in \Gamma\} \to \mathbb{R}^+ . \qquad (4)$$

For example $c(\gamma) = \mathbb{E}_\gamma(s(\beta))$ considers the expected value of the individual costs over the population, and $c(\gamma) = \mathbb{P}_{P_\gamma}(s(\beta) \geq \varepsilon) = \mathbb{P}_{P_\gamma}(\beta \notin V^{\mathrm{avg}})$ considers the percentage of cells whose behavior deviates from the reference by more than a user-defined threshold ε (cf. **Fig. 1B**); this percentage depends on the specific population distribution P_γ, and therefore on the population parameter γ.

Again, the population design problem can in principle be solved directly to yield $\gamma^* = \mathrm{argmin}_\gamma c(\gamma)$. Here, we again relax this problem and seek to identify the *population viable space* $V^{\mathrm{pop}} = \{\gamma \in \Gamma \mid c(\gamma) \leq \delta\}$ to account for additional uncertainties, where δ is again a user-defined parameter. In particular for design objectives such as requiring a minimal fraction of cells with 'acceptable' behavior that will have multiple optima, the population viable space also yields equivalent design alternatives.
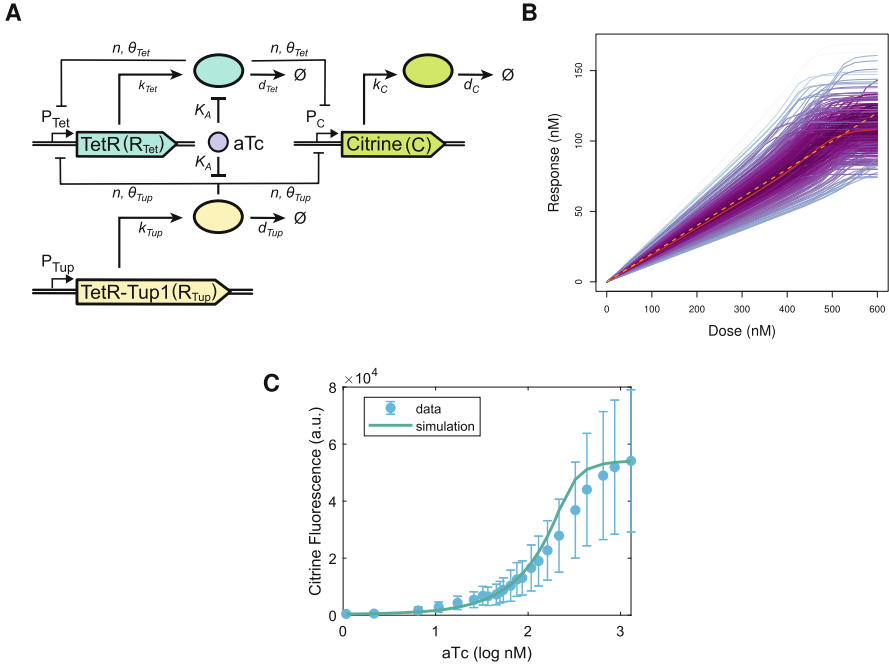
Fig. 2. Well-tempered controller (WTC) circuit. (A) Schematic representation of the circuit structure and its parametrization. Rectangles: genes with associated promoters; ellipses: proteins (corresponding color); bold lines with arrows: molecular reactions; normal lines with bar heads: regulatory interactions for inactivation. **(B)** Simulated dose-response curves of a population of cells for a given population parameter γ with a coefficient of variation $CV \approx 10\%$. Red line: median response; blue to purple lines: responses of individual cells colored by cost: the lower the cost, the darker the color; dashed orange line: reference linear dose-response curve, used to compute the individual cost. **(C)** Experimental and simulated aTc dose-response curve for the WTC. Blue: mean (circles) and standard deviation (error bars) of experimental data obtained by flow cytometry; green line: simulation results for the estimated parameter values in Table 1. Additionally, we used estimated values $d_C = 0.0031$ min^{-1}, $d_{Tet} = 0.005$ min^{-1}, $\theta_{Tet} = 1.2$ nM, and $\theta_{Tup} = 10^{-4}$ nM. To match the model output (Citrine concentration) to fluorescence (a.u.), we determined a scaling factor as in [10]. (Color figure online)

3 Case Study: Design of a Transcriptional Controller

3.1 Overview

To demonstrate the framework, we use a transcriptional controller termed well-tempered controller (WTC) that was experimentally designed by Azizoglu *et al.*

[1]. In the WTC (Fig. 2**A**), expression of the fluorescent protein Citrine—or of any gene of interest—is regulated by constitutively expressed TetR-Tup1 and by autorepressed TetR. Anhydrotetracycline (aTc) can bind to both TetR and TetR-Tup1, thereby inactivating their ability to repress gene expression.

Experimentally, it was shown that cell-to-cell variability in the expression of Citrine is reduced through the introduction of the TetR-mediated negative feedback. At the same time, the dose-response curve—obtained by adding different amounts of the inducer molecule aTc—was tuned to approach an ideal linear dose-response, corresponding to high Input Dynamic Range (IDR) and high Output Dynamic Range (ODR) [12] (Fig. 2**B**).

Given that we already know the final network structure of the WTC, we aim to use our computational framework to determine the acceptable characteristics of the distribution of circuit parameters in a population of cells, namely their mean and covariance, such that a large proportion of cells in the population will display a dose-response curve close to an ideal reference curve. Notably, we wish to establish whether our framework can identify the relevance of the feedback mechanism in the context of a population of cells.

3.2   Individual Model

We first formulated an ODE model to describe the behavior of the WTC circuit (see Fig. 2**A**). It involves the concentration of the input molecule aTc (a)—which can be added to the cell culture—and three states for the total concentrations of the repressor TetR (R_{Tet}), the repressor TetR-Tup1 (R_{Tup}) and the fluorescent protein Citrine (C):

$$\frac{\mathrm{d}R_{Tet}}{\mathrm{d}t} = \frac{k_{Tet}}{1 + \left(\frac{f \cdot R_{Tet}}{\theta_{Tet}}\right)^n + \left(\frac{f \cdot R_{Tup}}{\theta_{Tup}}\right)^n} - d_{Tet} \cdot R_{Tet} \tag{5}$$

$$\frac{\mathrm{d}R_{Tup}}{\mathrm{d}t} = k_{Tup} - d_{Tup} \cdot R_{Tup} \tag{6}$$

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \frac{k_C}{1 + \left(\frac{f \cdot R_{Tet}}{\theta_{Tet}}\right)^n + \left(\frac{f \cdot R_{Tup}}{\theta_{Tup}}\right)^n} - d_C \cdot C. \tag{7}$$

Parameters k_{Tet}, k_{Tup} and k_C are maximal expression constants that capture both transcription and translation to keep the model simple. Parameters d_{Tet}, d_{Tup} and d_C are the degradation constants.

For TetR and Citrine production we added a control term representing a Hill function that depends on the active concentrations of the repressors TetR and TetR-Tup1. Active TetR and TetR-Tup1 molecules are those that are not bound to the inducer aTc. Assuming rapid equilibrium for the binding of aTc to TetR and TetR-Tup1 (as in Lormeau *et al.* [10]), the fraction of active TetR and TetR-Tup1 (f) is given by:

$$f = \frac{1}{2} - \frac{1 + K_a a - \sqrt{(1 + K_a(R_{Tet} + R_{Tup} - a))^2 + 4K_a a}}{2K_a(R_{Tet} + R_{Tup})}. \tag{8}$$

Experimental data showed that TetR and TetR-Tup1 have different repression efficiencies [1], represented by θ in the model. We therefore decided to model the action of the two repressors on their controlled genes as an 'OR'-gate. This means that we are not taking into account that the repressors might bind to the same DNA sequences. In contrast, we do not expect a difference in Hill coefficient (n) or affinity (K_a) to aTc between TetR and TetR-Tup1.

3.3 Population Model

To simplify computations, we fixed the means of 6 out of 10 parameters of the ODE model (see Table 1). To obtain these values, we estimated the model parameters using data from Azizoglu *et al.* [1] and additional data on the WTC's biological parts. As shown in Fig. 2**C**, the parametrized WTC model captures the experimental dose-response curve.

The four remaining parameters ($d_{Tet}, d_C, \theta_{Tet}$, and θ_{Tup}) are the protein degradation constants, and the effective concentrations relative to the repression (including feedback) mechanisms. We fixed the mean value of d_{Tup} because this parameter is not identifiable together with θ_{Tup} using only steady-state information. If d_{Tup} were to be sampled along with θ_{Tup}, the strong negative correlation of these two parameters would not have any biological meaning. For the same reason, we fixed production constants and only allowed degradation constants to vary.

Regarding variances, only the production constants (k_{Tet}, k_{Tup}, k_C) and degradation constants (d_{Tet}, d_{Tup}, d_C) were assumed to display cell-to-cell variability. Without data on the variance of these parameters, we assumed that they all follow a log-normal distribution (to ensure positivity) based on the same variance σ^2 of the underlying Normal distribution. This implies that all the parameter distributions have the same coefficient of variation $CV = \sqrt{e^{\sigma^2} - 1}$. Since $\sigma \leq 0.1$ for our data, we use the approximation $CV \approx \sigma$ to simplify our analysis slightly.

3.4 Design Problem

Reference Dose-Response Curve. Our objective for the behavior of individual cells endowed with the WTC is a linear dose-response curve over an IDR of $[0\,\text{nM}, 600\,\text{nM}]$ for aTc with a desired ODR of $[0\,\text{nM}, 120\,\text{nM}]$. We encode a dose-response curve as a reference behavior. It takes the aTc concentration a as a constant input $u(t) \equiv a$, and yields a constant response $D^{\text{ref}}(a) \equiv D^{\text{ref}}(\tau; a)$ for all τ. We encode the high-IDR, high-ODR objective by defining $(a, D^{\text{ref}}(a))$ to be the straight line between $(0\,\text{nM}, 0\,\text{nM})$ and $(600\,\text{nM}, 120\,\text{nM})$.

Individual Cost. To quantify the deviation between an individual cell's behavior and the reference curve, we use the individual cost from Eq. 2 based on the dose-response curve $(a, D(\tau; \beta, a))$, where cell i has individual parameter set $\beta_i = (k_{Tet}^{(i)},\ k_{Tup}^{(i)},\ k_C^{(i)},\ d_{Tup}^{(i)},\ n,\ K_a,\ d_{Tet}^{(i)},\ d_C^{(i)},\ \theta_{Tet},\ \theta_{Tup})$, and $D(\tau; \beta, a) \equiv D(\beta, a)$ is the steady-state ($t \to \infty$) response to aTc concentration a. In our

Table 1. Parameter specifications for the WTC model. Parameters $k_{Tet}, k_{Tup}, k_C,$ and d_{Tup} are cell-to-cell variable but their mean is fixed to the indicated value.

Fixed parameters				
Name	Description	Units	Fixed value	Cell-specific
k_{Tet}	Max production rate of TetR	$\text{nM} \cdot \text{min}^{-1}$	1.115	Yes
k_{Tup}	Max production rate of TetR-Tup1	$\text{nM} \cdot \text{min}^{-1}$	0.7919	Yes
k_C	Max production rate of Citrine	$\text{nM} \cdot \text{min}^{-1}$	0.8395	Yes
d_{Tup}	Degradation constant of TetR-Tup1	min^{-1}	1.2745	Yes
n	Hill coefficient for promoter repression by TetR and TetR-Tup1	$(-)$	1.5656	No
K_a	Association constant for TetR and TetR-Tup1 binding to aTc	nM^{-1}	144.37	No
Sampled parameters				
Name	**Description**	**Units**	**Bounds**	**Explored in log space**	**Cell-specific**
d_{Tet}	Degradation constant of TetR	min^{-1}	$[10^{-5}\ 10]$	Yes	Yes
d_C	Degradation constant of Citrine	min^{-1}	$[10^{-10}\ 10]$	Yes	Yes
θ_{Tet}	Repression coefficient TetR	nM	$[10^{-6}\ 10^6]$	Yes	No
θ_{Tup}	Repression coefficient TetR-Tup1	nM	$[10^{-6}\ 20]$	Yes	No

implementation, the individual cost function is calculated via a discrete version of the L_2-norm based on N aTc input doses $\mathcal{U} = \{a_1, \ldots, a_N\}$, regularly spaced between 0 and 600 nM:

$$s(\beta) = \sqrt{\frac{1}{N} \sum_{k=1}^{N} \left(D(\beta, a_k) - D^{\text{ref}}(a_k)\right)^2} . \qquad (9)$$

We consider an individual cell's dose-response acceptable if $s(\beta) \leq \varepsilon$; the corresponding parameters β constitute the viable space. For our analysis, we use $\varepsilon = 5\,\text{nM}$ and $\varepsilon = 2\,\text{nM}$, which represent approximately 5% and 2% of the ODR we wish to achieve, respectively.

Population Cost. For our population design, we consider the percentage of individual cells in a population with parameter γ that fulfill the criterion Eq. 9 as our population cost function:

$$c(\gamma) = \mathbb{P}_{P_\gamma}(s(\beta) \geq \varepsilon) . \tag{10}$$

We define the population viable space as those γ that yield at least 80% individual cells with behavior sufficiently close to the reference and $c(\gamma) \leq 20\%$.

We estimate the population cost by drawing individual parameter sets β_i from the distribution P_γ and by determining the proportion of sampled parameter sets that yield acceptable individual costs.

Sampling in Parameter Spaces. We sampled from both the individual parameter space and the population parameter space, according to the individual cost s and the population cost c, respectively. We used an adaptive version of the Metropolis-Hastings algorithm [6] in both cases, implemented in the R [15] package 'fmcmc' [20], with pseudo-likelihoods based on individual cost and population cost. The package 'deSolve' [19] was used to solve the ODE model, with derivatives computed in C code. We defined the pseudo-likelihood for the individual parameter space as:

$$l(\beta) = \mathbb{1}(s(\beta) \leq \epsilon) \tag{11}$$

with $\epsilon \in \{5\,\text{nM}, 2\,\text{nM}\}$, therefore sampling uniformly the viable region $V^{\text{avg}} = \{\beta \in B \mid s(\beta) \leq \varepsilon\}$. The pseudo-likelihood for the population parameter space was:

$$L(\gamma) = \mathbb{1}(c(\gamma) \leq \delta) \tag{12}$$

with $\delta = 0.2$. We then obtain uniformly distributed samples from the population viable space $V^{\text{pop}} = \{\gamma \in \Gamma \mid c(\gamma) \leq \delta\}$. Note that, as $c(\gamma)$ depends on the value of ε (Eq. 10), $L(\gamma)$ and the associated population viable space will also depend on its value.

To compute this population pseudo-likelihood, however, we need to approximate $c(\gamma)$, as it is the functional of a distribution (in this case study, a probability). For each value of γ, 300 individual parameters were drawn randomly from the underlying log-normal distribution P_γ. For each individual parameter vector, we computed the individual cost s and approximated $c(\gamma)$ as the fraction of samples with individual costs above the corresponding threshold $\epsilon \in \{5\,\text{nM}, 2\,\text{nM}\}$. Note that we are interested in the resulting distribution of the individual costs and not in describing P_γ. Thus, even though we consider 6 cell-to-cell variable parameters, a sample size of 300 proved sufficient to reliably represent this distribution of individual costs as the underlying distance measure between a constant reference and the output of an ODE model is sufficiently smooth. An illustration is given in Fig. 2**B**, where 300 individual dose-response curves from a population distribution with high coefficient of variation cover the graph sufficiently.

The log-normal population distribution for our example allows us to reduce the required amount of random sampling and to provide more consistent results for the approximation of the population pseudo-likelihood. Note that we can

reconstruct the mean vector $\mu \in \mathbb{R}^6$ and the 6×6 covariance matrix C of the underlying multivariate Normal distribution from the population parameter γ. We therefore once generated 300 samples S_i from the standard multivariate Normal distribution $N(0, I)$ in $\mathbb{R}^6$. For each value of γ, we constructed the corresponding samples of the individual parameters as $\beta_i = \mu + C^{1/2} \cdot S_i$, where $C^{1/2}$ is the lower triangular matrix from a Cholesky decomposition of C. This ensures that repeated calls to our approximation of the population cost function with the same population parameter γ yields the same cost and requires only a single sample of size 300. On a standard laptop with Intel i7 processor, we obtained $\approx$900 samples from the population space per hour, corresponding to $\approx 2.7 \cdot 10^5$ samples from the individual parameter space.

Example. To illustrate the interplay between the individual and the population level in our design problem, Fig. 2**B** shows an example of the dose-response relationship of the WTC model for a population of cells. The NLME formulation takes into account the variance in parameters, that is, cell-to-cell variability. Here, although the median response is close enough to the ideal response, approximately 83% of the response curves are not within the acceptable range due to variance in the individual parameters. This leads to a population cost of $\approx$0.83, given an individual cost threshold of $\varepsilon = 5$ nM (corresponding to approximately 5% deviation from the reference curve). The example illustrates a key difference between traditional design assuming an 'average' cell and population design. If the design objective were to achieve a median response close to the reference, the example would be a valid solution, although the vast majority of individual cells would not comply with the design objective.

3.5   Sampling the Individual Parameters

Figure 3 shows the results of sampling the individual parameter space according to the value of the individual cost $s(\cdot)$. We first note that the protein degradation constant of Citrine, d_C, displays a substantially narrower marginal distribution than all other parameters. Citrine is the system response, and therefore this distribution shape is not surprising: with all other parameters kept identical, a change in d_C will directly impact the shape of the dose-response curve.

In the two-dimensional projections of the joint distribution over the individual viable space V^{avg}, the two parameters for protein degradation, d_{Tet} and d_C, are correlated, but mainly in the high-viability region. This indicates that either of the two parameters could be used to fine-tune the circuit.

Most importantly, the pattern of the projection across $(\theta_{Tet}, \theta_{Tup})$, which capture the strength of transcriptional repression, reveals insights into the relevance of negative feedback for WTC performance. Specifically, θ_{Tet} is the parameter for auto-repression, whereas θ_{Tup} is the parameter for constitutive repression. A smaller value of θ_{Tet} (resp. θ_{Tup}) means a stronger auto-repression (resp. constitutive repression). For the viability threshold of 5 nM used to define the viable space for the data shown in Fig. 3, most values for both θs are allowed, including high values that would effectively nullify the corresponding repressive
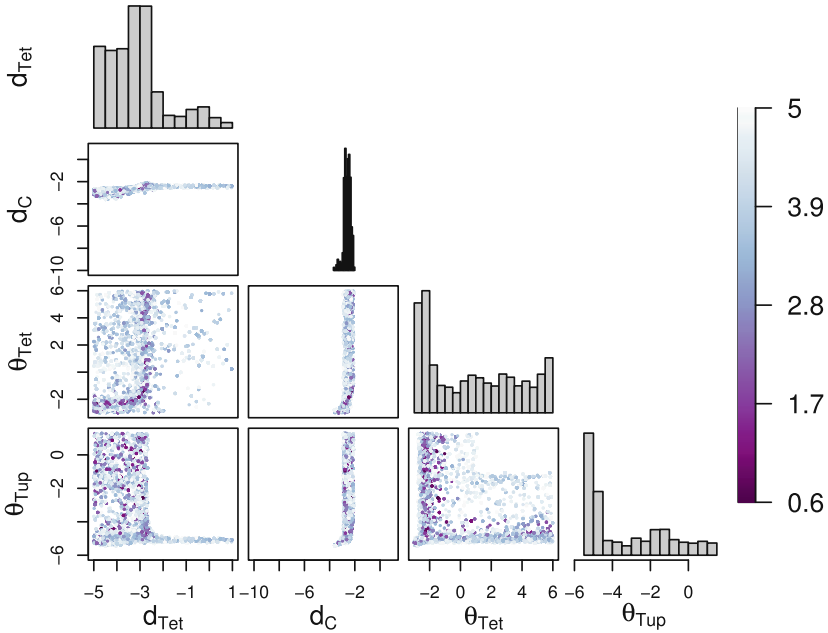
Fig. 3. Viable samples in the individual parameter space. Histograms show marginal distributions, and scatter plots samples in all two-dimensional projections of the parameter space. In the projections, samples are colored according to their individual cost from light blue to purple: a darker blue indicates a lower cost, and thus a higher consistency of the WTC dose-response with the reference curve for a given point. Only the parameters present in the plot were allowed to vary, all others were fixed to values specified in Table 1. Additionally, all parameters were sampled in $\log_{10}$-scale, and are displayed as such. (Color figure online)

effect. However, the upper right quadrant does not contain viable samples, indicating that at least one type of repression is needed for the circuit to achieve the desired behavior. Importantly, samples with lower values of the individual cost are located in the region of low θ_{Tet} (notice the color gradient). If we wish to achieve even closer correspondence of the WTC's dose-response with the reference curve for an individual cell (e.g., with an individual threshold $\varepsilon = 2\,\mathrm{nM}$), auto-repression becomes mandatory. Note as well that θ_{Tet} becomes strongly correlated with both degradation constants d_{Tet} and d_C, whenever auto-repression is strong. This is logical because auto-repression reduces the mean expression of TetR and Citrine, and should thus be compensated for by lower degradation constants to keep mean expressions in the desired range.

3.6   Sampling the Population Parameters

We next applied the population design framework described in Sect. 2 to the WTC model. Our aim is to obtain design guidelines for a reasonably good tran-
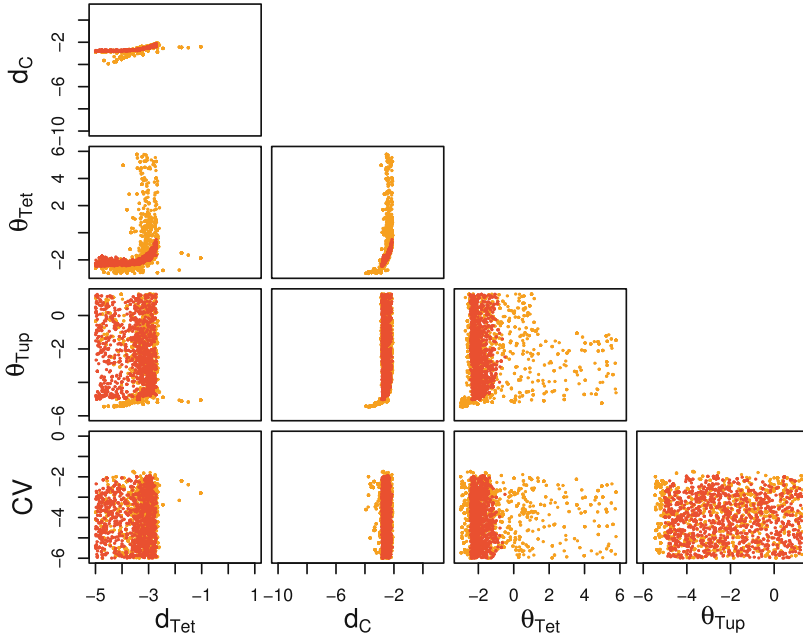
Fig. 4. Viable samples in the population parameter space. Samples in all two-dimensional projections of the parameter space; note that CV is the common coefficient of variation for all cell-to-cell variable population parameters. Orange dots: viable samples for the threshold on the individual cost $\varepsilon = 5\,\mathrm{nM}$; red dots: viable samples for $\varepsilon = 2\,\mathrm{nM}$. All parameters are in $\log_{10}$-scale.(Color figure online)

scriptional controller with low cell-to-cell variability in the steady-state dose-response, which we encode via the population cost Eq. 10 with a threshold of $\delta = 20\%$. The resulting samples according to Eq. 10 are shown in Fig. 4 for the two individual cost thresholds $\varepsilon = 5\,\mathrm{nM}$ (orange) and $\varepsilon = 2\,\mathrm{nM}$ (red). For both values of ε, we ran the Markov Chain Monte-Carlo (MCMC) chain twice from two different starting points. This explains the apparent density differences between regions, particularly visible in the planes (d_{Tet}, θ_{Tup}) and (d_{Tet}, CV).

Compared to the individual parameter samples (Fig. 3), we observe a clear upper bound of about $10^{-2.5}\mathrm{min}^{-1}$ for the population mean of d_{Tet}, which needs to be considered in the population design of the transcriptional controller.

Moreover, we find two distinct 'modes' of parameter combinations that lead to the desired population behavior, clearly visible in the (d_{Tet}, θ_{Tet}) panel of Fig. 4: if the average degradation constant d_{Tet} is large enough, this process alone ensures a level of $TetR$ compatible with the desired output and the auto-repression with θ_{Tet} can be chosen almost arbitrarily. Conversely, a low degradation constant requires strong auto-repression to achieve the population behavior, and thus low values for θ_{Tet}. These two modes are connected via a region with strong correlation between these parameters, indicating that both param-

eters need to be tuned simultaneously to achieve the population behavior in this region. In contrast, a strong auto-repression cannot compensate for a low degradation of Citrine (parameter d_C), while sufficiently high degradation of Citrine does not require tuning the auto-repression constants, as seen in panels (d_C, θ_{Tet}) and (d_C, θ_{Tup}).

To generate the data in Fig. 4, we allowed the coefficient of variation (CV, which is multiplicative in linear space) to vary up to a value of one. However, viable samples are essentially all below 0.02, pointing to this value as a possible maximum for the admissible cell-to-cell variability for reaching the design objective under the model's assumptions. For future studies, it is, hence, of interest to experimentally quantify the cell-to-cell variability of the parameters, and check the results against our inferred value. Note, however, that higher coefficients of variation would be allowed in the presence of negative correlations between parameters. In the plane (θ_{Tet}, CV) exists also a slightly decreasing slope for the case $\varepsilon = 5\,\text{nM}$: when the value of θ_{Tet} increases, leading to weaker auto-repression, the maximum admissible value for the coefficient of variation decreases. Indeed, the maximum CV for all samples has a value of $\approx 1.8\%$, whereas the maximum CV for the samples fulfilling the condition $\theta_{Tet} > 10^4\,\text{nM}$ is only $\approx 0.45\%$. This indicates that auto-repression can help compensate for cell-to-cell variability.

Regarding the repression parameters, θ_{Tet} and θ_{Tup}, we observe what could be expected from the individual samples: for $\varepsilon = 5\,\text{nM}$, the pattern of the projection of the samples over the plane $(\theta_{Tet}, \theta_{Tup})$ is very similar, if not exactly identical, to the one observed in Fig. 3. When the individual threshold is decreased to $\varepsilon = 2\,\text{nM}$, the viable region is reduced to low values of θ_{Tet}, indicating that auto-repression becomes necessary to achieve the design objective. Just as in the individual parameter space, θ_{Tet} strongly correlates with both degradation constants for low θ_{Tet}, *i.e.* for strong auto-repression. This applies particularly for $\varepsilon = 2\,\text{nM}$. If auto-repression is mandatory in a circuit, as here, particular attention should be given to tuning repression constants and degradation constants together.

Finally, we assumed that neither of the two repression parameters θ_{Tet} and θ_{Tup} displays cell-to-cell variability because the corresponding (microscopic) binding affinities are related to protein and DNA sequences that should be identical in each cell of an isogenic population. To assess the impact of this assumption, we performed an analogous sampling where the two parameters were assumed to vary from cell to cell just as the production and degradation constants; this yielded results very similar to Fig. 4 (data not shown).

4   Discussion

Nearly all current methods for synthetic circuit design assume an 'average' cell that needs to be optimized to fulfill the design objectives, potentially by considering parameter variations to achieve robustness of the biological implementation [10]. Stochastic design frameworks that account for cell-to-cell variability due

to intrinsic noise with low molecule copy numbers are beginning to emerge, but computational complexity currently limits them to small networks, steady-state, and homogeneous model parameters in a cell population [18]. Here, we therefore proposed population design via NLMEs as an alternative to both approaches. We argue that it has the potential to bring information about cell-to-cell variability to synthetic biological design in realistic settings, and to help infer the impact of said variability on the system of interest.

Our case study considers a problem synthetic circuit designers often face, namely to tune their system in order to reduce cell-to-cell variability [1]. For the WTC, the population sampling highlighted the importance of fine-tuning jointly the degradation constant of TetR and its auto-repression constant to achieve low cell-to-cell variability—the parameters could assume a wider range of values to achieve mere individual cell viability. Feedback mechanisms were necessary in both cases, at least under our assumption of a common variance parameter. This indicates that constitutive repression, and even more so auto-repression, are useful to linearize dose-response curves of individual cells. While constitutive repression had no impact on cell-to-cell variability, auto-repression could increase the admissible CV from $\approx 0.45\%$ to $\approx 1.8\%$. However, we could not achieve higher values of the CV, most likely because variability reduction is directly linked to repression strength: increasing repression would decrease cell-to-cell variability as well as mean expression of the repressed component. To weaken or eliminate this link between mean and variability, one may need to consider more complex topologies [3]. Note also that we limited our analysis to a small number of dimensions. Future studies could include more parameters or allow all variance parameters to be sampled independently. With independently sampled variances, it would be particularly interesting to see how autorepression affects (presumably relaxes) variance constraints across the network.

One limitation of our study (and an impediment to the extended analysis of the WTC) is the sampling technique we used. MCMC sampling does not scale well with dimensions, but one could use dedicated methods for sampling in higher dimensional spaces [22] instead. We also noted a tendency of the MCMC chain to get stuck in some parameter regions for population sampling, thus requiring multiple starting points to explore the whole space; this was not needed for the individual parameter space. However, keeping in mind that the number of variance parameters (including correlations) grows quadratically with the number of individual parameters, it is likely that one will not be able to tune the variance of each parameter individually. As a possible strategy, one could fix the covariance matrix to an experimentally determined one, for example, by using well-established NLME inference approaches [4,8] to obtain a parametrization of the cell-to-cell variability of biological parts. Other (not mutually exclusive) alternatives include the use of approximation methods for the individual cost [16] and of small sets sampling techniques such as the sigma-point approximation [9]. A different approach could be to replace exact MCMC sampling by approximate methods. For example, variational inference can be much faster than MCMC

and still provide accurate results, provided that the correlation structure of the likelihood is properly accounted for [5].

For the present case study, we explored the population parameter space of a network topology we knew should work for some parameter values. In the broader context of synthetic biology, a working, simple topology that has the potential to achieve the design objective is not necessarily known. In many cases, one may want to explore different topologies and select the one that performs best while still being simple enough. To achieve this goal while taking into account cell-to-cell variability, we propose to apply the method described by Lormeau *et al.* [10] to the objective function defined at the population level. Briefly, the algorithm will explore a number of possible topologies by simplifying an initial (complex) starting network, removing its edges. The viability (existence of parameters making the network viable) of each network is assessed. One can then choose robust networks according to the size of the viable region, for instance. The case study presented here only aimed at providing a first insight into the relevance of sampling from the population viable space, but we did not sample the population viable space for multiple topologies. However, our findings for the WTC on the importance of feedback mechanisms (to achieve the design objective, without impacting cell-to-cell variability) and of fine-tuning TetR degradation (to reduce cell-to-cell variability) indicate that the concept is promising.

Overall, the population design framework could then be used to recommend network structures, together with their parameter values, that are best suited to fulfill a design objective incorporating cell-to-cell variability. Such an approach could also help exploring situations where cell-to-cell variability and a given distribution over behaviors of cells in a population is desirable. One example is bet-hedging in bacterial populations, where non-genetic variability across a population increases the chances of survival in the face of antibiotics [13].

5   Conclusion

We propose a general framework we call population design that aims to help biologists interested in synthetic circuit design to account for cell-to-cell variability via ODE-based NLMEs. We implemented a simple version of the concept and demonstrated its usefulness for a transcriptional controller in an a posteriori case study. The current implementation is restricted to small models with few parameters. We hope to augment it with advanced numerical methods and extend it to the problem of topology design. In perspective, this could enable the rational design of synthetic gene circuits that induce prescribed (distributions of) behaviors at the population level, and thereby allow to exploit cell-to-cell variability for novel applications.

References

1. Azizoğlu, A., Brent, R., Rudolf, F.: A precisely adjustable, variation-suppressed eukaryotic transcriptional controller to enable genetic discovery. bioRxiv p. 2019.12.12.874461 (2020). https://doi.org/10.7554/eLife.69549

2. Barnes, C.P., Silk, D., Sheng, X., Stumpf, M.P.H.: Bayesian design of synthetic biological systems. Proc. Nat. Acad. Sci. **108**(37), 15190–15195 (2011). https://doi.org/10.1073/pnas.1017972108

3. Bonny, A.R., Fonseca, J.P., Park, J.E., El-Samad, H.: Orthogonal control of mean and variability of endogenous genes in a human cell line. Nature Commun. **12**(1), 1–9 (2021). https://doi.org/10.1038/s41467-020-20467-8

4. Dharmarajan, L., Kaltenbach, H.M., Rudolf, F., Stelling, J.: A simple and flexible computational framework for inferring sources of heterogeneity from single-cell dynamics. Cell Syst. **8**(1), 15–26.e11 (2019). https://doi.org/10.1016/j.cels.2018.12.007

5. Ghosh, S., Birrell, P., De Angelis, D.: Variational inference for nonlinear ordinary differential equations. In: Banerjee, A., Fukumizu, K. (eds.) Proceedings of The 24th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 130, pp. 2719–2727. PMLR (2021). http://proceedings.mlr.press/v130/ghosh21b.html

6. Haario, H., Saksman, E., Tamminen, J.: An adaptive Metropolis algorithm. Bernoulli **7**(2), 223–242 (2001). https://doi.org/10.2307/3318737

7. Karamasioti, E., Lormeau, C., Stelling, J.: Computational design of biological circuits: putting parts into context. Mol. Syst. Des. Eng. **2**(4), 410–421 (2017). https://doi.org/10.1039/C7ME00032D

8. Lavielle, M.: Mixed effects models for the population approach: models, tasks, methods, and tools. CPT: Pharmacometrics Syst. Pharmacol. **4**(1), (2015). https://doi.org/10.1002/psp4.10

9. Loos, C., Moeller, K., Fröhlich, F., Hucho, T., Hasenauer, J.: A hierarchical, data-driven approach to modeling single-cell populations predicts latent causes of cell-to-cell variability. Cell Syst. **6**(5), 593–603.e13 (2018). https://doi.org/10.1016/j.cels.2018.04.008

10. Lormeau, C., Rudolf, F., Stelling, J.: A rationally engineered decoder of transient intracellular signals. Nature Commun. **12**(1), 1886 (2021). https://doi.org/10.1038/s41467-021-22190-4

11. Lormeau, C., Rybiński, M., Stelling, J.: Multi-objective design of synthetic biological circuits. IFAC-PapersOnLine **50**(1), 9871–9876 (2017). https://doi.org/10.1016/j.ifacol.2017.08.1601

12. Mannan, A.A., Liu, D., Zhang, F., Oyarzún, D.A.: Fundamental design principles for transcription-factor-based metabolite biosensors. ACS Synth. Biol. **6**(10), 1851–1859 (2017). https://doi.org/10.1021/acssynbio.7b00172

13. Martín, P.V., Muñoz, M.A., Pigolotti, S.: Bet-hedging strategies in expanding populations. PLOS Comput. Biol. **15**(4), e1006529 (2019). https://doi.org/10.1371/journal.pcbi.1006529

14. Nielsen, A.A.K., et al.: Genetic circuit design automation. Science **352**, 6281 (2016). https://doi.org/10.1126/science.aac7341

15. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2021). https://www.R-project.org/

16. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, Cambridge, Mass, Adaptive Computation and Machine Learning (2006)

17. Ryan, E.G., Drovandi, C.C., McGree, J.M., Pettitt, A.N.: A review of modern computational algorithms for Bayesian optimal design. Int. Stat. Rev. **84**(1), 128–154 (2016). https://doi.org/10.1111/insr.12107
18. Sakurai, Y., Hori, Y.: Optimization-based synthesis of stochastic biocircuits with statistical specifications. J. Royal Soc. Interface **15**,(2018). https://doi.org/10.1098/rsif.2017.0709
19. Soetaert, K., Petzoldt, T., Setzer, R.W.: Solving differential equations in R: Package deSolve. J. Stat. Soft. **33**(9), 1–25 (2010). https://doi.org/10.18637/jss.v033.i09
20. Vega Yon, G., Marjoram, P.: fmcmc: A friendly MCMC framework. J. Open Source Softw. **4**(39), (2019). https://doi.org/10.21105/joss.01427
21. Voigt, C.A.: Synthetic biology 2020–2030: six commercially-available products that are changing our world. Nature Commun. **11**, 6379 (2020). https://doi.org/10.1038/s41467-020-20122-2
22. Zamora-Sillero, E., Hafner, M., Ibig, A., Stelling, J., Wagner, A.: Efficient characterization of high-dimensional parameter spaces for systems biology. BMC Syst. Biol. **5**(1), 142 (2011). https://doi.org/10.1186/1752-0509-5-142
23. Češka, M., Dannenberg, F., Paoletti, N., Kwiatkowska, M., Brim, L.: Precise parameter synthesis for stochastic biochemical systems. Acta Informatica **54**(6), 589–623 (2017). https://doi.org/10.1007/s00236-016-0265-2