RESEARCH

Open Access

Efficient design of synthetic gene circuits under cell-to-cell variability

Baptiste Turpin[1], Eline Y. Bijman[1], Hans-Michael Kaltenbach[1] and Jörg Stelling[1*]

*Correspondence:
joerg.stelling@bsse.ethz.ch

[1] Department of Biosystems
Science and Engineering
(D-BSSE) and SIB Swiss Institute
of Bioinformatics, ETH Zurich,
4058 Basel, Switzerland

Abstract

Background: 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.

Results: Here, we present a computational framework and an efficient algorithm to guide the design of synthetic biological circuits while accounting for cell-to-cell variability explicitly. Our design method integrates a Non-linear Mixed-Effects (NLME) framework into a Markov Chain Monte-Carlo (MCMC) 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.

Conclusion: 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 supporting design objectives that specify the desired behavior of cell populations.

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

Background

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 [1]. 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 [2].

Computer-aided design helps identifying suitable network structures (topologies) as well as biological parts for their implementation to reach a given design 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. 1A, B) [3–5].

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 [6]) 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 [7]). One can account for uncertainty in ODE-based design, for example, via measures of robustness that quantify parameter uncertainty [4]. It is also possible to tackle cell-to-cell variability with stochastic models, where temporal logic specifications are written as Continuous Stochastic Logic (CSL) [8]. 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 MCMC based sampling approaches for ODE-based design [4] to the NLME (Non-linear Mixed-Effects) models framework [9]. 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
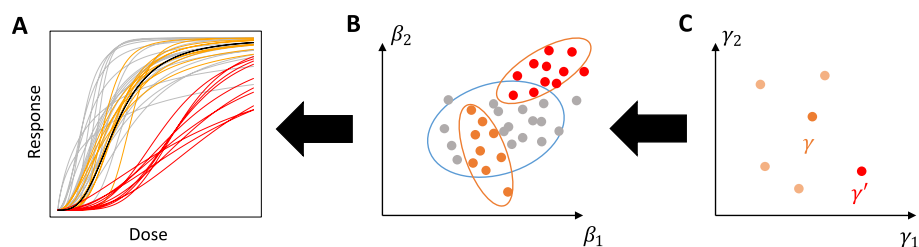
Fig. 1 Cell behaviors relate to parameters at the individual and population level. **A** Dose-response relationships for single cells (lines) drawn from two distinct populations (red and orange) as well as other cells not belonging to any population (gray). The design objective for individual cells is 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**)

Turpin *et al. BMC Bioinformatics*   (2023) 24:460

Page 3 of 26

level (see Fig. 1B, C). This allows a designer to impose cell-to-cell variability constraints on synthetic networks. Additionally, we propose an efficient MCMC algorithm with parallel tempering (PT) [10] to scale our approach to larger population design problems. We demonstrate the approach with the a posteriori analysis of a recently developed transcriptional controller [11], a class of circuits that is often designed to minimize cell-to-cell variability.

Results

Population design framework

For any individual cell, we assume that 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 vector $\beta = (\alpha, x^0) \in B$, where B is a bounded set.

We first consider the *individual 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 vector $\beta \in B$ and any input $u \in \mathcal{U}$, where $\mathcal{U}$ is a finite set of relevant inputs. The *reference behavior* $D^{\text{ref}} : \mathbb{R} \times \mathcal{U} \to \mathbb{R}$ 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*

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

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

In principle, the individual cell design problem could be solved directly to identify the optimal individual cell parameter $\beta^* = \text{argmin}_\beta s(\beta)$. However, additional uncertainties arise due to unmodelled system components and from combining previously characterized biological parts into a circuit [12]. 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* $V^{\text{ind}} = \{\beta \in B \mid s(\beta) \leq \varepsilon\}$ of all parameters that fulfill this

criterion. An output of the individual cell design problem is then a description of V^{ind} rather than a single parameter.

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*

$$\beta_i \sim P_\gamma \tag{3}$$

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. We assume that the distribution P_γ admits a probability density function (p.d.f.) $p_\gamma(\beta)$ for all $\gamma \in \Gamma$.

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

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

For example, $c(\gamma) = \mathbb{E}_\gamma(s(\beta))$ considers the expected value of the individual costs over the population, and $c(\gamma) = \mathbb{P}_{P_\gamma}(s(\beta) \geq \varepsilon) = \mathbb{P}_{P_\gamma}(\beta \notin V^{\mathrm{ind}})$ 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.

To sample the viable spaces, we previously applied a naive MCMC-based algorithm to a low-dimensional design problem, but it proved to be slow and had a poor mixing [13]. It would be increasingly difficult and time-consuming to apply it to larger problems involving more parameters. To overcome these computational limitations for the important case of controlling the percentage of sufficiently well-behaved cells, here we propose and implement an alternative sampling algorithm that we call Stochastic Likelihood Markov Chain Monte-Carlo (SLMCMC). Its general idea is to sample jointly from the population parameter space and the individual parameter space, and then to discard individual parameters (see Methods for details).

Design problem for a transcriptional controller

To demonstrate the framework, we use a transcriptional controller termed well-tempered controller (WTC) that was experimentally designed by Azizoglu et al. [11]. In the WTC (Fig. 2A), expression of the fluorescent protein Citrine—or of any gene of
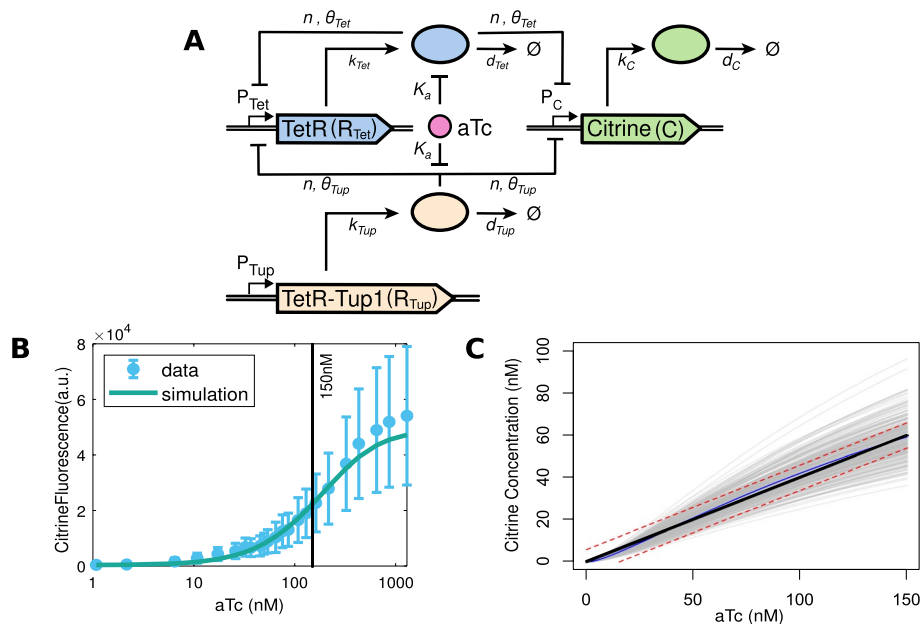
Fig. 2 Well-tempered controller (WTC) circuit. **A** Schematic representation of the circuit structure and its parametrization. Rectangles: genes with associated promoters; ellipses: proteins (corresponding color); lines with arrows: molecular reactions; lines with bar heads: regulatory interactions for inactivation. **B** 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 $d_C = 0.006$ min^{-1}, $d_{Tet} = 0.0087$ min^{-1}, $\theta_{Tet} = 0.006$ nM, and $\theta_{Tup} = 0.5 \cdot 10^{-4}$ nM. To match the model output (Citrine concentration) to fluorescence (a.u.), we determined a scaling factor as in [4]. **C** Simulated dose-response curves of a population of cells for a given population parameter γ with a coefficient of variation $CV \approx 10\%$. Parameter means were the estimated parameters from (**B**). Blue line: mean response; gray lines: responses of individual cells; black line: reference linear dose-response curve; red lines: reference curve $\pm \varepsilon = 6$ nM; the individual cost threshold; an individual trajectory is viable if it lies within the red curves on average. Note that aTc is in linear scale

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) [14] (Fig. 2B).

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.

We first formulated an ODE model to describe the behavior of the WTC circuit (see Fig. 2A) for an individual ('average') cell. The model 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, the repressor TetR-Tup1, and the fluorescent

protein Citrine (see Methods for details). We estimated the model's 10 parameters using data from Azizoglu et al. [11] using least-squares non-linear optimization (see Table 1). Note that compared to our earlier study [13], we make the model biologically more realistic by accounting for dilution due to cell growth through the degradation constants. As shown in Fig. 2B, the parametrized WTC model captures the experimental dose-response curve well. In particular, the model can give a response close to linear in the input range shown in Fig. 2C. Here, the parameterized model serves only to generate a realistic reference behavior; therefore we did not investigate model aspects such as identifiability.

To define the design problem, we encode this observed behavior as the reference behavior. Specifically, our objective for the behavior of individual cells endowed with the WTC is a linear dose-response curve over an IDR of [0 nM, 150 nM] for aTc with a desired ODR of [0 nM, 60 nM]. We define our individual cost $s(\beta)$ (Eq. 2) as the L_2 distance between an individual cell's response and the reference dose-response curve. Parameters β that fulfill $s(\beta) \leq \varepsilon$ constitute the individual viable space, and we consider different values for the threshold ε (see Methods).

Table 1 Parameter specifications for the WTC model

Name	Description	Units	Fixed value	Cell-specific
Fixed parameters				
k_{Tet}	Max production rate of TetR	nM min^{-1}	1.12	Yes
k_{Tup}	Max production rate of TetR-Tup1	nM min^{-1}	0.79	Yes
k_C	Max production rate of Citrine	nM min^{-1}	0.84	Yes
d_{Tup}	Degradation constant of TetR-Tup1	min^{-1}	1.28	Yes
n	Hill coefficient for promoter repression by TetR	(–)	1.57	No
K_a	Association constant for TetR and TetR-Tup1 binding to aTc	nM^{-1}	144.37	No
U_{sc}^{nM}	Scaling constant: nM per tdh3 producing unit (see [4])	nM · unit^{-1}	0.76	No

Name	Description	Units	Bounds	Explored in log space	Cell-specific
Sampled parameters					
d_{Tet}	Degradation constant of TetR	min^{-1}	$[3 \cdot 10^{-3} \, 0.1]$	Yes	Yes
d_C	Degradation constant of Citrine	min^{-1}	$[3 \cdot 10^{-3} \, 0.1]$	Yes	Yes
θ_{Tet}	Repression coefficient TetR	nM	$[10^{-3} \, 10^6]$	Yes	No
θ_{Tup}	Repression coefficient TetR-Tup1	nM	$[10^{-6} \, 20]$	Yes	No

Parameters k_{Tet}, k_{Tup}, k_C, and d_{Tup} are cell-to-cell variable but their means are fixed to the indicated values. *S. cerevisiae*'s typical growth rate in YPD medium is $0.0077 \, \text{min}^{-1}$ [4]. To account for dilution and potential cell-to-cell variability, lower bounds for all degradation constants were fixed to $3e - 3 \, \text{min}^{-1}$.

For our population design, we consider the percentage of individual cells in a population with parameter γ that fulfill the criterion for acceptable dose-response as our population cost function:

$$c(\gamma) = \mathbb{P}_{P_\gamma}(s(\beta) \geq \varepsilon) . \tag{5}$$

We define the population viable space as those γ that yield at least 80% individual cells with behavior sufficiently close to the reference; correspondingly, $c(\gamma) \leq 0.2$.

To illustrate the interplay between the individual and the population level in our design problem, Fig. 2C 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 mean response is close enough to the ideal response, many response curves are not within the acceptable range due to variance in the individual parameters. Specifically, we obtain a population cost of $c(\gamma) \approx 0.3$, given an individual cost threshold of $\varepsilon = 6$ nM that corresponds to approximately 10% deviation from the reference curve. The example illustrates a key difference between traditional design and population design. The traditional design considers the mean response of the population, which is very close to the reference curve, although a significant fraction of individual cells might not comply with the design objective. In contrast, the population design explicitly considers variability and rejects design options if the proportion of non-compliant cells is too high.

Individual-cell design

We were first interested in identifying the relevance of the feedback mechanism for single-cell responses, focusing on individual costs and (viable) parameter spaces. To simplify the computations as well as the analysis of relationships between parameters, we sampled a four-dimensional parameter space after fixing the means of 6 out of 10 parameters of the ODE model (see Table 1). 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. This set of parameters also allows to explore limit cases where parts of the network are removed: A high repression constant (θ_{Tet} or θ_{Tup}) is equivalent to removing the corresponding repressive effect; fixing d_{Tet} or d_C to a large value effectively removes the production of TetR or Citrine.

For sampling, we used a naive, adaptive version of the Metropolis-Hastings algorithm [15] with a pseudo-likelihood based on individual cost (see Methods for details). We selected a threshold of $\varepsilon = 6$ nM for the individual cost, corresponding to approximately 10% of the target ODR.

In Fig. 3, 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.

The two-dimensional projections of the joint distribution over the individual viable space V^{ind} exhibit a correlation between the two parameters for protein degradation, d_{Tet} and d_C, mainly in the region of low individual cost $s(\beta)$. This indicates that either of the two degradation constants could be used to fine-tune the circuit.

The two parameters θ_{Tet} and θ_{Tup} capture the strengths of auto-repression and constitutive repression, respectively. Higher values correspond to weaker repression, but do not exclude other vital roles for TetR (resp. TetR-Tup1) in the circuit. Values for the full range of θ_{Tet} and θ_{Tup} are viable individually. However, the two parameters cannot take high values simultaneously, indicating that effective repression of at least one type is needed for the circuit to achieve the desired behavior.

A high degradation constant d_{Tet} might allow us to remove TetR from the circuit design and rely on constitutive repression alone. However, looking at the (d_{Tet}, θ_{Tup}) projection reveals that this option requires precisely controlling constitutive repression by keeping θ_{Tup} in a very narrow range. This is unlikely to be feasible in a biological implementation—keeping TetR in the circuit therefore seems advisable even without a repressive role.

For low values of θ_{Tet} and θ_{Tup}, these two parameters are also correlated with the degradation constant of Citrine, d_C, because stronger repression needs to be compensated by slower degradation to allow mean expression of Citrine in the desired range. Citrine mean expression seems less affected by lower repression, where this correlation vanishes.

Considering viable parameter values jointly thus provides insights into parameter restrictions and correlations. It also identifies robustness with respect to parameter perturbations that can be exploited for circuit design.

Population design

We next applied the complete population design framework to the WTC model to obtain design guidelines for a reasonably good transcriptional controller with low cell-to-cell variability in the steady-state dose-response. Specifically, we set a population threshold of $\delta = 0.2$, requiring at least 80% of the cells in a population to meet the individual design goal.

Table 1 summarizes the parameter specifications. Assuming log-normal distributions, we sample only variances (k_{Tet}, k_{Tup}, k_C, and d_{Tup}), only means (θ_{Tet} and θ_{Tup}), or both (d_{Tet} and d_C). While variances are unlikely to be tunable in practice, sampling cell-to-cell variability of parameters allows us to identify maximum admissible values compatible with the design objectives and hence to select suitable biological parts (of known and fixed variance) for the circuit.

As a main extension to our prior work [13], we here introduce the SLMCMC algorithm for sampling in population design. The previous, naive approach is computationally inefficient because substantial sampling of individual parameters is required to impose the hard threshold δ on the population cost in Eq. 6. We reasoned that we might gain efficiency by relaxing the hard constraint during sampling, focusing on sampling viable population parameters with sufficient probability, and checking the hard δ-constraint *a posteriori*. For details on the algorithm, see Methods.

To test the algorithm, we first assume an identical coefficient of variation $CV = \sqrt{e^{\sigma^2} - 1}$ for all parameters that are cell-to-cell variable. The covariance matrix of the underlying Normal distribution is then $\sigma^2 \cdot I$. Since $\sigma \leq 0.1$ in our viable samples, we used the approximation $CV \approx \sigma$ to simplify our analysis slightly.

Figure 4A shows the resulting population parameter distributions for the SLMCMC algorithm using the two individual cost thresholds $\varepsilon = 6\,\text{nM}$ and $\varepsilon = 3\,\text{nM}$. They are very similar to the ones obtained with the naive approach (Additional file 2: Fig. S1), indicating that the SLMCMC approach finds parameters yielding the desired behavior at the population level, although it samples from a different distribution initially (see Methods for details).

For the mean parameters, the patterns are also very similar to those obtained for individual samples (Fig. 3). This is not surprising because we allowed for low values for the variance σ in the exploration; for low σ, essentially all population members behave like the (viable) mean parameters.

The data in Fig. 4A provides new insights, first, because of the two individual thresholds ε employed. A reduced threshold makes the correlation between d_{Tet} and d_C stronger, highlighting the importance of tuning the two parameters jointly to achieve a stricter objective. In line with the individual design results, the joint distribution of repression parameters θ_{Tet} and θ_{Tup} shows that either type of repression needs to be strong to achieve the design objective. In addition, for lower ε, the region of low θ_{Tet} and high θ_{Tup} recedes to higher θ_{Tet}, indicating that simultaneous strong auto-repression and weak constitutive repression can be detrimental for stricter objectives.

The second new insight derives from sampling parameter variances in population design. Although we allowed a CV up to one, Viable samples are essentially all below 7% for $\varepsilon = 6$ nM and 2.4% for $\varepsilon = 3$ nM. These values could be maxima for the admissible cell-to-cell variability for reaching the design objective. 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 CVs would be allowed in the presence of negative correlations between parameters.

Here, we sampled the *CV* to identify the maximum admissible variability. To assess the impact of a fixed (high) variability on the viable space, we then fixed $CV = 7\%$. In addition, because non-linearities should make differences between individual and population design more pronounced [16], we increased the Hill coefficient for TetR-Tup1 ($n_{Tup} = 5$). This makes the roles of constitutive repression and auto-repression more distinct. As shown in Fig. 4B, the population viable space becomes restricted to parts of the space where individual cells would have a very low individual cost. As a result, some parameter combinations are no longer viable for populations including variability. For instance, d_{Tet} and d_C have much higher correlation in the population viable space, and will need well-coordinated fine-tuning.

Next, we abandoned the assumption of equal variances and considered a diagonal covariance matrix with diagonal entries σ_j^2, which increases the parameter space to 10 dimensions. The resulting distribution for the mean parameters is very similar to the one obtained for the scalar covariance matrix (Additional file 3: Fig. S2). However, viable variances differ considerably (Fig. 5A): $\sigma_{d_{Tup}}$ and $\sigma_{k_{Tup}}$ can rise to CVs as high as 1. Thus, TetR-Tup1 degradation and production constants can vary substantially in the population without much impact on the performance. Yet, σ_{d_C}, $\sigma_{d_{Tet}}$, σ_{k_C}, and $\sigma_{k_{Tet}}$ are constrained to values below 10%, indicating that narrow distributions for the corresponding constants are critical for adequate circuit performance.

Turpin *et al. BMC Bioinformatics*   (2023) 24:460

Page 10 of 26

Table 2 Sampling efficiency in the population parameter space

Criterion	MCMC	SLMCMC				
Samples	20,000	200,000				
Covariance	Scalar	Scalar				Diagonal
κ	–	3	{3,5}	{3,7}	{3,9}	{3,7}
Run-time (s)	23,755	900	2639	3624	3713	3413
Checking-time (s)	0	615	721	719	725	746
Rejected samples (%)	0	4	1.33	1.33	0.5	2.33
minESS	157	1505	3011	3345	2981	1487
Time/sample (s)	1.188	0.0076	0.017	0.020	0.022	0.021
Time/minESS (s)	151.31	1.01	1.12	1.19	1.49	2.80
Sampling efficiency (%)	0.79	0.75	1.51	1.67	1.49	0.74

Runs were performed on a standard laptop with Intel i7 processor for the population sampling problem with $\varepsilon = 6nM$. Run-time is the sampling time. Checking-time is the time to naively compute population costs for 600 sampled populations (unnecessary for the naive approach). Populations with cost above the threshold account for the rejected samples. minESS is the minimum of the individual ESSs obtained across multiple dimensions. Sampling efficiency is the ratio between minESS and the actual number of samples drawn. Ratios involving time are computed by summing over run-time and checking-time
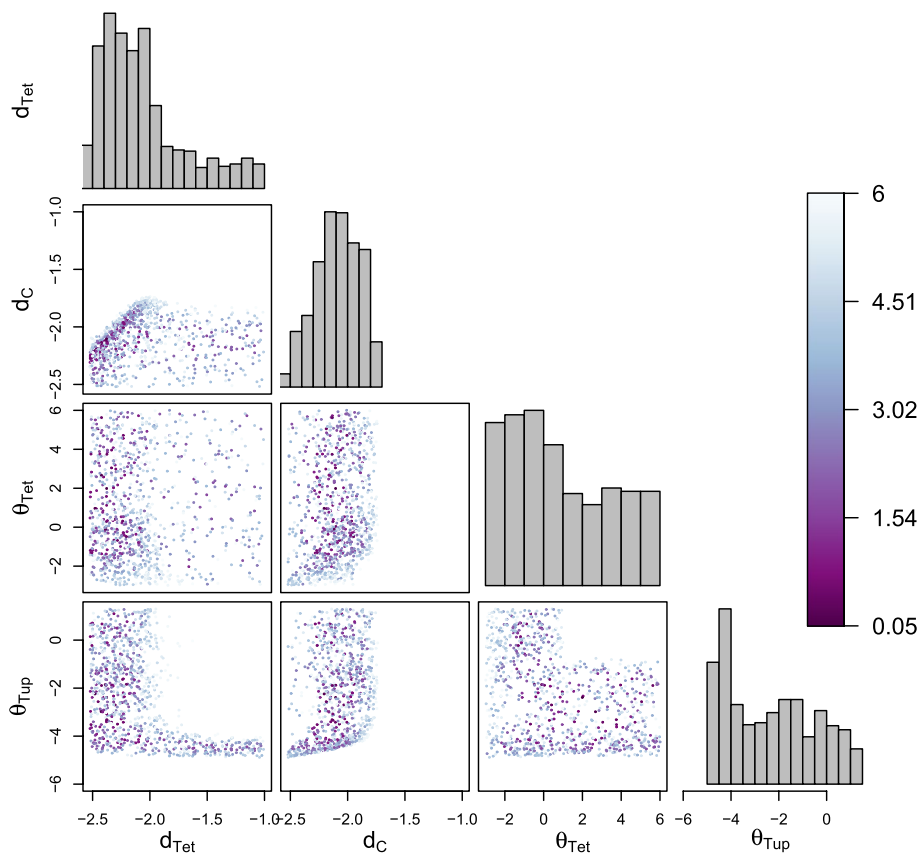
Fig. 3 Viable samples in the individual parameter space. Histograms show marginal distributions, and scatter plots samples in all two-dimensional projections of the parameter space. In the projections, samples are colored according to their individual cost from light blue to purple: a darker blue indicates 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

Tight control of d_C and k_C is unsurprising as these parameters are directly related to Citrine, the system's output. The importance of controlling TetR concentration is less obvious, as $\sigma_{k_{Tet}}$ and $\sigma_{d_{Tet}}$ can rise to higher values for $\varepsilon = 6$nM. However, high values of $\sigma_{k_{Tet}}$ are always associated with high d_{Tet} and a narrow distribution of θ_{Tup} (Fig. 5B, first two panels). This corresponds to the limiting case of very strong constitutive repression combined with very low amounts of TetR, in which case the variance of TetR is unimportant. Indeed, a high value of d_{Tet} forces a lower variability on TetR-Tup1 parameters k_{Tup} and d_{Tup} (Fig. 5B, last two panels). If TetR is removed from the system (very high d_{Tet} value), TetR-Tup1 must show little cell-to-cell variability and a very precise repression strength.

Considering parameter-specific variances in the population design problem thus provides additional insights because parameters with low variance require tighter control for the circuit to work.

Sampling efficiency

We quantified the efficiency of the naive approach and of our new SLMCMC approach with and without parallel tempering by recording the run times for generating samples and by computing the minimal Effective Sample Size (minESS) as the minimum Effective Sample Size (ESS) over all dimensions (Table 2).

For the naive approach, we generated 20,000 samples for the population design problem with $\varepsilon = 6$ nM. We checked convergence of the chain using the Gelman–Rubin criterion [17], indicating that several thousand samples were needed for convergence. This resulted in a total run-time of more than 23,000 s and an effective sample size of 157. This corresponds to more than 1 s per sample and a sampling efficiency of 0.79%. Importantly, though, the naive approach ensures that all sampled population parameters fulfill the population design criterion of at least 80% of the cells meeting the individual target.

In contrast, our SLMCMC approach does not guarantee the hard population threshold; it requires post-processing of the sampled parameters to check if the target was indeed met. We distinguish these two stages in our calculations. The SLMCMC does not require estimating the population cost in each step—a very costly computation. This allows generating samples much more efficiently, and we therefore generated 200,000 samples. Using a parameter $\kappa = 3$, a scalar covariance matrix, and no parallel tempering, these samples are generated in about 900 s. We argue that it is not necessary to check the population design criterion for each sample, and use a statistical approach instead: we check the criterion only for a randomly picked subsample of 600 out of the 200,000 samples. This provides a reliable estimate of the percentage of population parameters that

(See figure on next page.)

Fig. 4 Comparison of individual and population viable spaces. **A** Samples in all two-dimensional projections of the population parameter space obtained with the SLMCMC algorithm for a scalar covariance matrix (less than 1.4% of samples are not viable). *CV* is the common coefficient of variation for all cell-to-cell variable population parameters. Orange (red) dots: viable samples for the threshold on the individual cost $\varepsilon = 6$ nM ($\varepsilon = 3$ nM); black dots: populations with cost $c(\gamma) > 0.2$. All parameters are in $\log_{10}$-scale. **B** WTC model with increased cooperativity of repression. Orange (blue) dots: Individual samples with cost above (below) $\varepsilon = 3$ nM, individual threshold $\varepsilon = 6$ nM. Red dots: Population samples with fixed $CV = 7\%$, population threshold $\delta = 0.2$
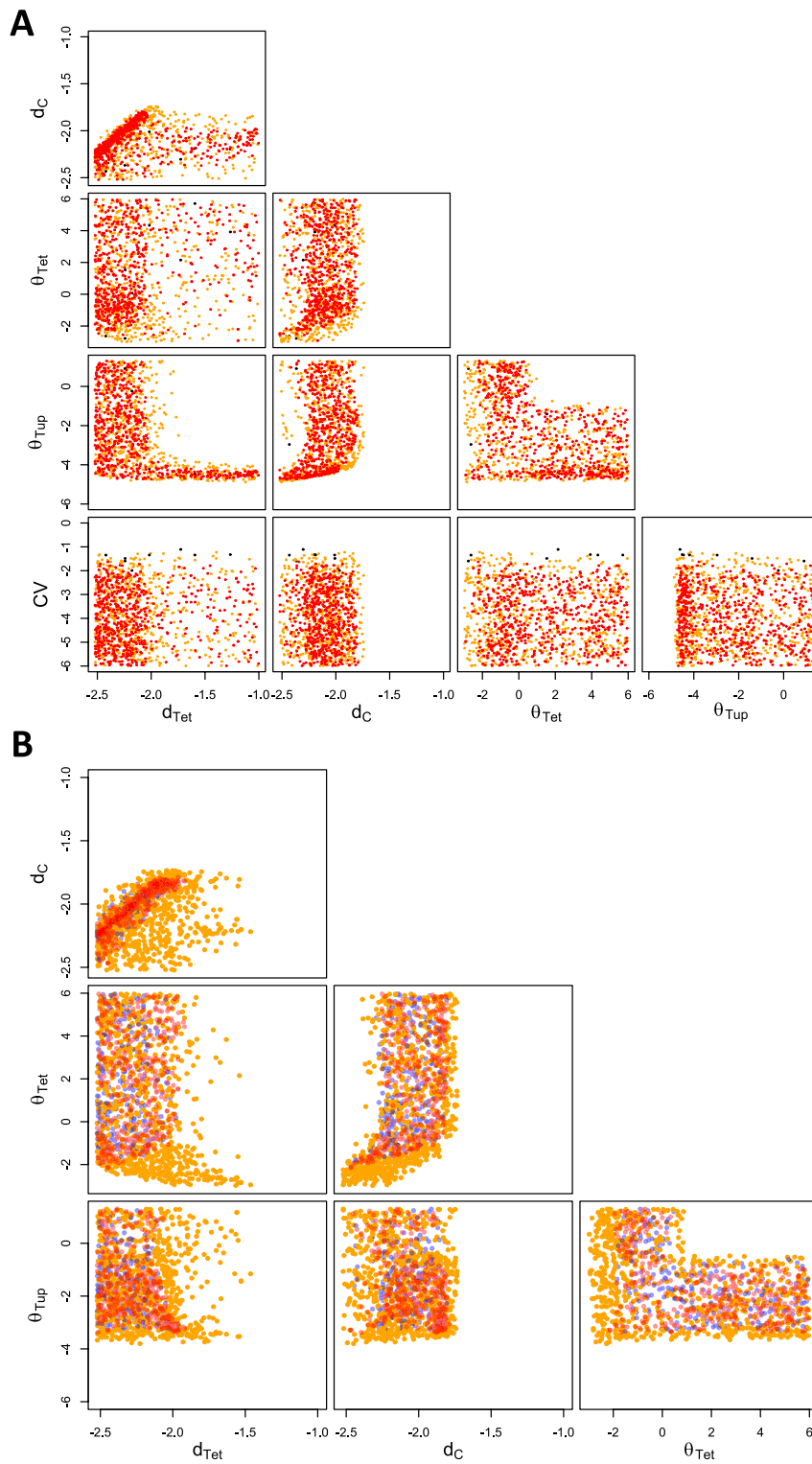
Fig. 4 (See legend on previous page.)

do not meet the hard threshold. We found that only an estimated 4% of the samples did not meet the population design criterion. The minimal ESS is comparable to the naive approach, as is the sampling efficiency.

A

B
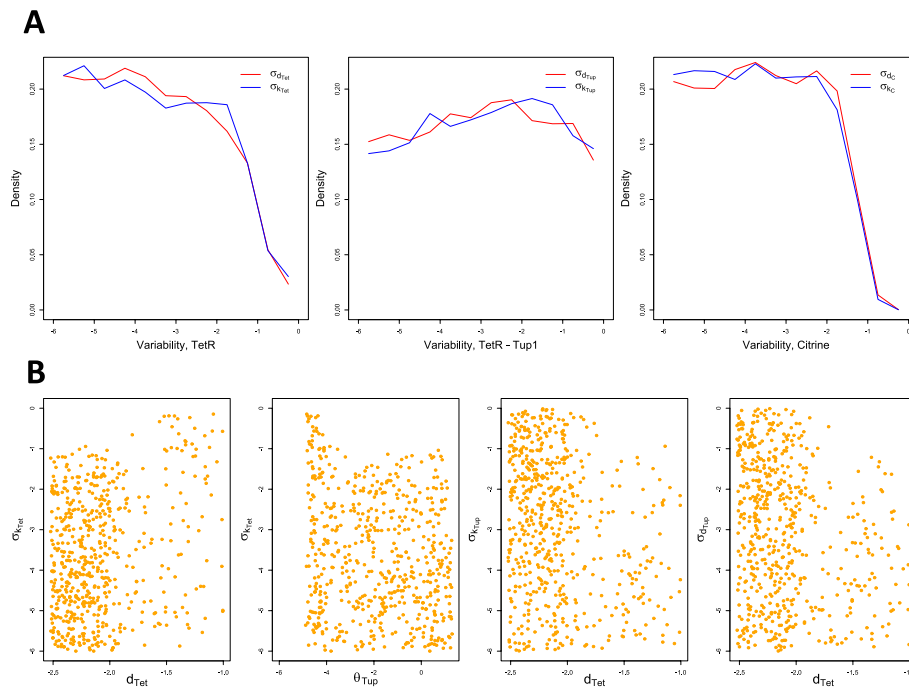
Fig. 5 Distributions of sampled population parameters, diagonal covariance matrix. Individual threshold is set to $\varepsilon = 6\,\text{nM}$. For the full distributions, see Additional file 3: Fig. S2. **A** Marginal distributions of samples for cell-to-cell variability in the population parameter space. **B** Viable samples in selected two-dimensional projections of the population parameter space

To evaluate the impact of parallel tempering on performance, we chose chains with $\kappa_1 = 3$ and $\kappa_2 = 5, 7, 9$. The run-time for generating 200,000 samples increased to about 2.8–4-fold compared to the non-tempered approach, but the percentage of samples not meeting the target dropped to 0.5–1.3%. The minimal ESS roughly doubled, leading to an overall sampling efficiency of about 1.5%, double the efficiency of the naive and the non-tempered approaches. We again selected 600 samples at random to check the population design criterion, which added about 700 s to the overall computation time.

Finally, we selected the most efficient parallel tempering setting with $\kappa_1 = 3$ and $\kappa_2 = 7$ and used a diagonal instead of a scalar covariance matrix. The overall run-time is comparable to the previous settings, but the fraction or rejected samples increased slightly to about 2.3%. More importantly, several variance parameters need to be sampled in this setting, and the increased problem dimension leads to a minimum ESS of 1487, about half of the other PT-based scenarios. The sampling efficiency is then comparable to the naive approach and the non-tempered SLMCMC for scalar covariance matrices.

Overall, our SLMCMC algorithm with parallel tempering outperforms the naive approach independent of the specific setting for the inverse temperatures. It generates about two orders of magnitude more effective samples per time, shows an order of magnitude increase in minimal ESS and a massive reduction in overall run-time.

Extended population design

Finally, we asked if population design could help to extend the WTC's input dynamic range. Specifically, starting from the current input dynamic range with maximum

maxIDR $= 150$ nM, we aimed to extend it to 300 nM, 450 nM, and 600 nM, respectively, while restricting the slope of the dose-response curve to the current value. We used the SLMCMC approach with an individual threshold of $\varepsilon = 6$ nM and a diagonal covariance matrix to sample the population parameter space. The top-right of Fig. 6 presents examples of populations for each of the three design objectives, alongside the corresponding reference curves.

Most population parameters are little affected by extending the IDR, but the maximal viable values for d_C and θ_{Tup} decrease considerably with increasing maxIDR (Fig. 6). The viable space for maxIDR $= 450$ nM is restricted to very low values of these two parameters, which limits options for implementation severely. However, auto-repression seems less important in this case, and θ_{Tet} becomes less restricted.

For variance parameters, patterns are similar to those observed previously, with the exception that parameters related to TetR-Tup1 become more constrained with increasing maxIDR. For high maxIDR, auto-repression does not help to achieve the design objective and the variances of k_{Tup} and d_{Tup} become independent of d_{Tet}. Meanwhile, high variances of d_{Tet}, k_{Tet} remain viable only for high d_{Tet}. This indicates that if TetR is
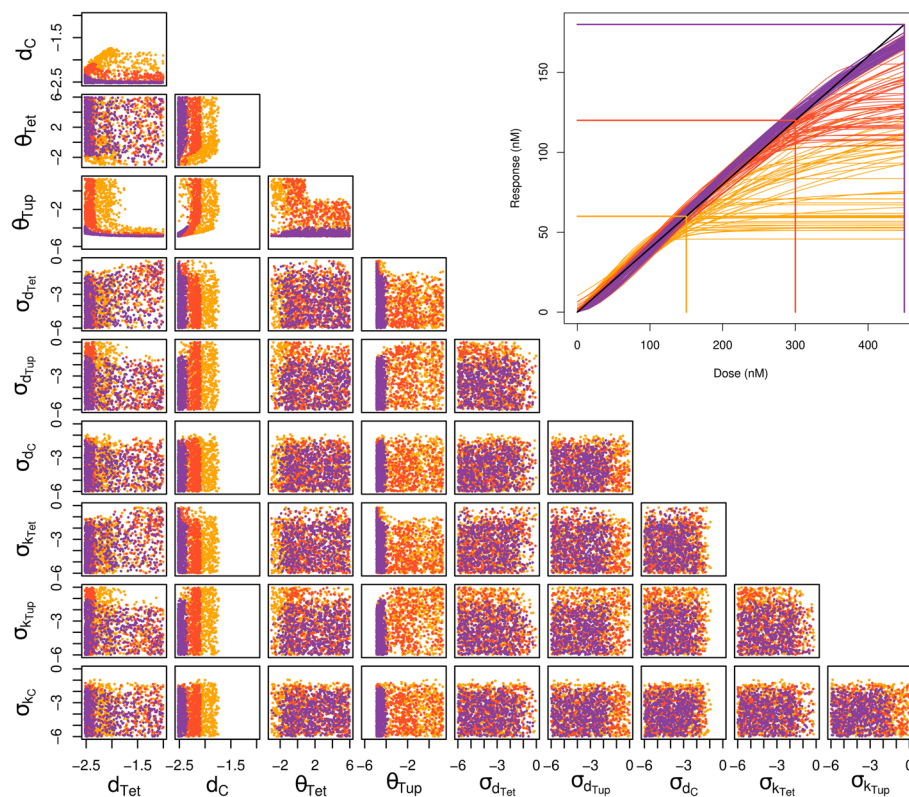
Fig. 6 Viable samples in the population parameter space for reference curves with extended IDR. Samples in all two-dimensional projections of the parameter space, obtained with the SLMCMC algorithm for individual threshold $\varepsilon = 6$ nM and diagonal covariance matrix. Orange dots: IDR up to 150 nM; red dots: IDR up to 300 nM; purple dots: IDR up to 450 nM nM. Non-viable population samples (i.e. $c(\gamma) > 0.2$) were removed prior to plotting. All parameters were sampled in $\log_{10}$-scale, and are displayed as such. Top-right corner: Corresponding viable population dose-response curves. Black line: reference curve up to 450 nM. Colored lines corresponding to sample colors: 50 population dose-response curves per reference curve, computed as in Fig. 2C from 50 samples of population parameters. Colored boxes delimit the portion of the black line used as a reference curve during population design

Turpin *et al. BMC Bioinformatics*   (2023) 24:460

Page 15 of 26

present in the system (low d_{Tet}), its variability should be controlled, as its binding with aTc may otherwise propagate variability to the fraction of bound TetR-Tup1 and further to Citrine.

The observed parameter restrictions become untenable for maxIDR = 600 nM and our method found no viable population parameter set for this case. Indeed, with a minimum individual cost above 8 nM, no individual cell can satisfy the individual cost requirement given the WTC model's parametrization and constraints on the parameter spaces.

Thus, our population design approach suggest that—and how—the WTC's input dynamic range can be extended, and it gives an indication of the limits to such an extension under the current circuit design (and parametrization).

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 [4]. 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. We additionally propose the SLMCMC algorithm as an efficient way to solve the population design problem *via* sampling.

Our case study considers a problem synthetic circuit designers often face, namely to tune their system in order to reduce cell-to-cell variability [11].

Repression mechanisms were necessary both in the individual and population cases. This indicates that constitutive repression and auto-repression are useful to linearize dose-response curves of individual cells. Interestingly, TetR seemed to be able to fulfill that role even when it does not enact auto-repression.

The population sampling with scalar covariance matrix highlighted 7% as a possible maximum admissible coefficient of variation, whereas our analysis with a diagonal covariance matrix showed that the variances of different parameters play different roles. For the WTC, parameters related to Citrine and TetR have to have low cell-to-cell variability, but TetR-Tup1 permits higher variability, as long as TetR is present in the system.

Constitutive repression increased the admissible CV from $\approx 4\%$ to $\approx 7\%$. However, we do not know if constitutive repression had a direct impact on cell-to-cell variability, or if it simply helped linearize the dose-response curve. To disentangle these possibilities, we would need to define a measure of variability reduction independent of the shape of the response, and to compare this measure for different repression strengths. Additionally, any variability reduction will be 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 circuit topologies [19, 20].

In view of enhancing the WTC's performance, a more restrictive design objective (larger IDR and ODR) could be obtained by tuning precisely the constitutive repression strength and the degradation constant of Citrine. Constitutive repression proved essential to increase IDR and ODR, in contrast to auto-repression. At least with the given model and parametrization, however, we do not predict that the highest IDR we tested can be achieved by the WTC.

Note that some of the analysis results for the WTC differ from our previously published results [13]. There, we used the same model, but lower bounds on degradation constants for a proof-of-concept. Additionally, we fixed the maximal simulated time to 1000 min, which is a reasonable time in an experimental setting. Here, we accounted for effective degradation constants closer to the experimentally observed dilution rate, leading to increased lower bounds and a different design objective. In addition, we adjusted simulation times for the system to reach steady-states. These changes in combination explain the differences between the present and the earlier study.

In terms of methods, our new sampling technique for population parameters, the SLMCMC algorithm, proved reliable and fast as it bypasses the need to compute the population cost. The added benefit of the parallel tempering approach is an increased mixing, especially when the target distribution is multimodal with poorly connected regions of high density. SLMCMC worked efficiently for our rather high-dimensional applications, but it is still an MCMC algorithm, and as such it becomes less efficient in higher dimensions. To reduce this limitation, SLMCMC could be combined with techniques such as evolutionary MCMC [21] to improve mixing. However, because 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 for large models. Instead, one could fix the covariance matrix to estimates obtained from experimental data, for example, by using well-established NLME inference approaches [7, 9]. A complementary approach would be to approximate the individual cost [22]. Other alternatives, which are not compatible with SLMCMC, include small sets sampling techniques such as the sigma-point approximation [23], or replacing exact MCMC sampling by approximate methods. For example, variational inference can be much faster than MCMC and still give accurate results, provided that the correlation structure of the likelihood is properly accounted for [24].

Here, 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 a method similar to that described by Lormeau et al. [4] 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. 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.

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 [25].

Conclusions

Our general framework for population design, along with the efficient SLMCMC algorithm, aim to help biologists interested in synthetic circuit design to account for cell-to-cell variability via ODE-based NLMEs. The a posteriori demonstration of its usefulness for a transcriptional controller shows scaling to a relevant problem size, although the implementation suffers from classic sampling inefficiency. In perspective, an extension to topology design 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.

Methods

Overview

Our main objective is a description of the population viable space V^{pop} that encompasses all solutions of the population design problem. As this space will typically not have an explicit description, we resort to MCMC methods to sample it. We propose three sampling strategies: a naive MCMC approach, a new method that exploits joint sampling of the individual and population parameter spaces which we call *Stochastic Likelihood MCMC (SLMCMC)*, and a modified version of SLMCMC that uses parallel tempering to increase sampling efficiency. We restrict attention to our WTC problem and use the proportion of viable individual cells as our population design criterion. The naive MCMC approach readily generalizes to other population objectives, but the SLMCMC approach is currently tailored to this specific objective.

Naive algorithm for population sampling

Our previously described naive sampling approach [13] is a straightforward implementation of the Metropolis-Hastings sampler. We directly use the population design criterion as our pseudo-likelihood:

$$L(\gamma) = \mathbb{1}(c(\gamma) \leq \delta). \tag{6}$$

The resulting samples from the population parameter space Γ will have a p.d.f. proportional to $L(\gamma)$, which constrains them to the population viable space. The function $L(\gamma)$ is not amenable to direct calculation because the population cost $c(\gamma)$ is not known explicitly. We therefore draw N random individual parameters β_i from the corresponding population distribution P_γ to approximate $c(\gamma)$ by $\hat{c}(\gamma) = \sum_{j=1}^{N} \mathbb{1}(s(\beta_j) \leq \varepsilon)/N$ and $L(\gamma)$ by $\hat{L}(\gamma) = \mathbb{1}(\hat{c}(\gamma) \leq \delta)$. We use $N = 300$ for our computations, which provided sufficient sample size for our problem. The full approach is given in Algorithm 1, where we use c and L instead of $\hat{c}$ and $\hat{L}$ for readability.

Algorithm 1 Naive algorithm for population sampling

Require: $\gamma_0 \in \Gamma$, $N > 0$
  Draw $\beta_1^{(0)}, \ldots, \beta_N^{(0)} \sim P_{\gamma_0}$ i.i.d.
  $c(\gamma) \leftarrow \sum_{j=1}^N \mathbb{1}(s(\beta_j^{(0)}) \leq \varepsilon) / N$
  $L_0 \leftarrow \mathbb{1}(c(\gamma) \leq \delta)$
Ensure: $L_0 > 0$
  for $i \in \{0, \ldots, N-1\}$ **do**
      Draw $\gamma' \sim q(\cdot|\gamma_i)$.
      Draw $\beta_1', \ldots, \beta_N' \sim P_{\gamma'}$ i.i.d.
      $c(\gamma') \leftarrow \sum_{j=1}^N \mathbb{1}(s(\beta_j') \leq \varepsilon) / N$
      $L' \leftarrow \mathbb{1}(c(\gamma') \leq \delta)$
      $\alpha(\gamma'|\gamma_i) \leftarrow \min \left(\dfrac{L'q(\gamma_i|\gamma')}{L_i q(\gamma'|\gamma_i)}, 1 \right)$
      Draw $v \sim \mathcal{U}([0,1])$
      if $v < \alpha(\gamma'|\gamma_i)$ **then**
          $\gamma_{i+1} \leftarrow \gamma'$
          $L_{i+1} \leftarrow L'$
      else
          $\gamma_{i+1} \leftarrow \gamma_i$
          $L_{i+1} \leftarrow L_i$
      end if
  end for

SLMCMC algorithm

To derive the algorithm, note that Eq. 6 can be rewritten as

$$L(\gamma) = \mathbb{1}\left(\mathbb{P}_{P_\gamma}(s(\beta) > \varepsilon) \leq \delta \right), \tag{7}$$

where $\beta \sim P_\gamma$ is considered a random variable. Removing the indicator and the population threshold, and taking the complement probability yields the objective

$$L(\gamma) = \mathbb{P}_{P_\gamma}(s(\beta) \leq \varepsilon) = 1 - c(\gamma) \tag{8}$$

which provides the desired relaxed version of Eq. 6.

   The central idea is then to extend this objective to the product space $\Gamma \times B$ of population and individual parameters:

$$L(\gamma, \beta) = \mathbb{1}(s(\beta) \leq \varepsilon) \cdot p_\gamma(\beta), \tag{9}$$

where p_γ is the density of the log-normal distribution P_γ. Evaluating Eq. 9 and sampling from the joint distribution is straightforward. The relaxed objective Eq. 8 is then the marginal

$$L(\gamma) = \int_B L(\gamma, \beta) d\beta, \tag{10}$$

MCMC sampling from $L(\gamma, \beta)$ followed by discarding the β component will provide γ samples distributed according to the objective (Eq. 8).

   In a last step, we introduce a constant exponent $\kappa \in \mathbb{N}_{>0}$ to concentrate the pseudo-likelihood around favorably high values, a standard technique used in Bayesian design [26], for example. This yields the objective

$$L(\gamma) = \mathbb{P}^{\kappa}_{P_{\gamma}}(s(\beta) \leq \varepsilon),\tag{11}$$

which we sample using the corresponding joint pseudo-likelihood

$$L(\gamma, \beta_1, \ldots, \beta_{\kappa}) = \prod_{j=1}^{\kappa} \mathbb{1}(s(\beta_j) \leq \varepsilon) \cdot p_{\gamma}(\beta_j).\tag{12}$$

Intuitively, this means that instead of a single random individual parameter yielding sufficiently low individual cost, we now require that κ independent parameters simultaneously yield low costs. The higher κ is, the lower the population costs of the sampled parameters will be on average and the more stringent the following sampling will be.

We propose Algorithm 2 to sample from the joint distribution (Eq. 12). We use a Metropolis-Hastings algorithm with the custom proposal distribution:

$$q(\gamma', \beta'_1, \ldots, \beta'_{\kappa} | \gamma, \beta_1, \ldots, \beta_{\kappa}) = q(\gamma' | \gamma) \prod_{j=1}^{\kappa} p_{\gamma'}(\beta'_j),\tag{13}$$

where $q(\gamma'|\gamma)$ is a user-defined proposal distribution, restricted to the population parameter. In other words, we first draw a proposal for the population parameter; this proposal depends on the last population parameter, but not on the last individual parameters. After that, we draw the proposal for the individual parameters directly as independent samples from $P_{\gamma'}$. This cancels the conditional probability terms in the acceptance ratio. We thus need to store $\prod_{j=1}^{\kappa} \mathbb{1}(s(\beta_j) \leq \varepsilon)$, but not the individual parameters.

Algorithm 2 SLMCMC algorithm for population sampling

Require: $\gamma_0 \in \Gamma, \kappa > 0, N > 0$
  Draw $\beta_1^{(0)}, \ldots, \beta_{\kappa}^{(0)} \sim P_{\gamma_0}$ i.i.d.
  $L_0 \leftarrow \prod_{j=1}^{\kappa} \mathbb{1}\left(s(\beta_j^{(0)}) \leq \varepsilon\right)$
Ensure: $L_0 > 0$
  for $i \in \{0, \ldots, N-1\}$ **do**
    Draw $\gamma' \sim q(\cdot|\gamma_i)$
    Draw $\beta'_1, \ldots, \beta'_{\kappa} \sim P_{\gamma'}$ i.i.d.
    $L' \leftarrow \prod_{j=1}^{\kappa} \mathbb{1}\left(s(\beta'_j) \leq \varepsilon\right)$
    $\alpha(\gamma'|\gamma_i) \leftarrow \min\left(\frac{L' q(\gamma_i|\gamma')}{L_i q(\gamma'|\gamma_i)}, 1\right)$
    Draw $v \sim \mathcal{U}([0,1])$
    if $v < \alpha(\gamma'|\gamma_i)$ **then**
      $\gamma_{i+1} \leftarrow \gamma'$
      $L_{i+1} \leftarrow L'$
    else
      $\gamma_{i+1} \leftarrow \gamma_i$
      $L_{i+1} \leftarrow L_i$
    end if
  end for

Two technical aspects are worth noting: first, L_0 and L' are products of indicator functions; we can stop sampling values β_i as soon as the first indicator function is zero to further reduce the computational burden. Second, sampled individual parameters β_i are already 'likely' as they are drawn from the population distribution with parameter γ and therefore follow the required conditional distribution. Hence, we do not need to compute the conditional probabilities directly. This would also allow us to

replace P_γ with more complex distributions that do not admit a p.d.f. in closed form, but from which it is reasonably easy to sample.

An important difference between standard Metropolis-Hastings and our variant is the computation of the likelihood ratio, which is stochastic in our algorithm (hence the name SLMCMC). An equivalent technique is known as likelihood-free sampling or subset simulation in the context of Bayesian models, where it is applied to sampling datasets rather than parameters [27, 28].

SLMCMC with parallel tempering

Our SLMCMC approach for sampling the viable population space is still an MCMC technique and suffers from the associated drawbacks, particularly poorly scaling with problem dimension and poor mixing for multimodal distributions. Poor scaling can be addressed using adaptive proposals [15, 29, 30], and we now describe a parallel tempering (PT) [31] extension to improve mixing.

Parallel tempering uses several parallel chains with limiting distributions forming a sequence from a 'flat' to the target distribution. A popular family of distributions uses an inverse temperature parameter $\lambda \in [0, 1]$ and limiting distributions of the form f^λ, where f is the desired target. Chains with higher temperature (lower λ) have 'flatter' distributions which makes it easier to explore the sampling space and avoid local modes. In addition, two chains with limiting distributions f and f' can 'swap' their current states x and x' with probability

$$\alpha_{\text{swap}}(f, f') = \frac{f(x')f'(x)}{f(x)f'(x')} .$$

$$(14)$$

Swapping states renders the individual chains non-Markovian, but the set of chains remains Markovian on the product space with the product distribution of the chains as its invariant distribution [10].

To apply the parallel tempering approach to SLMCMC, we use the number of individual parameters κ to provide the inverse temperature. Specifically, consider the pseudo-likelihood Eq. 11 with exponent κ_{max} as our sampling target. We specify a sequence of r chains whose targets are the pseudo-likelihoods Eq. 12 with different values $\kappa_1 < \kappa_2 < \cdots < \kappa_r = \kappa_{\text{max}}$ for the parameter κ. This generates a family of r distributions for parallel tempering with inverse temperatures $\kappa_i/\kappa_{\text{max}}$.

These chains have different state-spaces $\Gamma \times B^{\kappa_i}$, which prohibits direct application of the swapping procedure. We therefore extend the pseudo-likelihoods to the highest-dimensional state-space $\Gamma \times B^{\kappa_{\text{max}}}$ and write the likelihood for the κ_i chain as

$$L_{\kappa_i}(\gamma, \beta_1, \ldots, \beta_{\kappa_{max}}) = \prod_{j=1}^{\kappa_{max}} \mathbb{1}\left(s(\beta_j) \leq \varepsilon\right)^{\mathbb{1}(j \leq \kappa_i)} p_\gamma(\beta_j) .$$

$$(15)$$

This effectively sets the 'acceptance' of parameters β_j to one for $j > \kappa_i$. With this provision, and applying Eq. 14, the acceptance probability to swap between two chains with parameters $\kappa < \kappa'$ and samples β_i, β'_i, respectively, is

$$
\alpha_{\text{swap}}(\kappa,\kappa') = \min\left(\frac{\prod_{j=1}^{\kappa}\mathbb{1}(s(\beta_j')\leq\varepsilon)\prod_{j=1}^{\kappa'}\mathbb{1}(s(\beta_j)\leq\varepsilon)}{\prod_{j=1}^{\kappa}\mathbb{1}(s(\beta_j)\leq\varepsilon)\prod_{j=1}^{\kappa'}\mathbb{1}(s(\beta_j')\leq\varepsilon)},1\right)
$$

$$
= \min\left(\prod_{j=\kappa+1}^{\kappa'}\frac{\mathbb{1}(s(\beta_j)\leq\varepsilon)}{\mathbb{1}(s(\beta_j')\leq\varepsilon)},1\right)
$$

$$
= \prod_{j=\kappa+1}^{\kappa'}\mathbb{1}(s(\beta_j)\leq\varepsilon)
$$

because $\mathbb{1}(s(\beta_j')\leq\varepsilon)=1$ for $j>\kappa$ and the product cannot exceed one in any case.

Our SLMCMC approach with parallel tempering is given in Algorithm 3. Note that the required numbers of samples β_j depend on the tempering parameters κ_i: for two chains with $\kappa_1 < \kappa_2$, we need $\kappa_1 + \kappa_2$ samples of individual parameters to update the two chains, and additional $\kappa_2 - \kappa_1$ individual parameters for swapping, that is, $2\cdot\kappa_2$ sampled individual parameters in total.

Algorithm 3 Parallel-Tempering SLMCMC algorithm for population sampling

---

Require: $\left(\gamma_0^{(1)},\ldots,\gamma_0^{(r)}\right)\in\Gamma^r$, $(\kappa_1,\ldots,\kappa_r)\in(\mathbb{N}^*)^r$, $N>0$

for $\kappa\in\{\kappa_1,\ldots,\kappa_r\}$ **do**
    Initialize the chain associated with inverse temperature κ as in algorithm 2
end for
for $i\in\{0,\ldots,N-1\}$ **do**
    for $\kappa\in\{\kappa_1,\ldots,\kappa_r\}$ **do**
        Update the chain associated with inverse temperature κ as in algorithm 2
    end for
    Randomly select 2 adjacent chains
    Denote their respective inverse temperatures $\kappa<\kappa'$ and population states γ and γ'
    Draw $\beta_1,\ldots,\beta_{\kappa'-\kappa}\sim P_\gamma$ i.i.d.
    $w\leftarrow\prod_{j=1}^{\kappa'-\kappa}\mathbb{1}(s(\beta_j)\leq\varepsilon)$
    if $w=1$ **then**
        Swap states γ and γ'
    end if
end for

WTC model

We model the dynamics of the total concentrations of the repressor TetR (R_{Tet}), the repressor TetR-Tup1 (R_{Tup}) and the fluorescent protein Citrine (C) by:

$$
\frac{\mathrm{dR}_{\text{Tet}}}{\mathrm{dt}} = \frac{k_{Tet}}{1+\left(\frac{f\cdot R_{Tet}}{\theta_{Tet}}\right)^n+\left(\frac{f\cdot R_{Tup}}{\theta_{Tup}}\right)^n} - d_{Tet}\cdot R_{Tet} \tag{16}
$$

$$
\frac{\mathrm{dR}_{\text{Tup}}}{\mathrm{dt}} = k_{Tup} - d_{Tup}\cdot R_{Tup} \tag{17}
$$

$$
\frac{\mathrm{dC}}{\mathrm{dt}} = \frac{k_C}{1+\left(\frac{f\cdot R_{Tet}}{\theta_{Tet}}\right)^n+\left(\frac{f\cdot R_{Tup}}{\theta_{Tup}}\right)^n} - d_C\cdot C. \tag{18}
$$

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.

The two Hill functions represent control terms for TetR and Citrine production, respectively. They depend 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. [4]), the fraction of active TetR and TetR-Tup1 (f) is given by:

$$f = \frac{1}{2} - \frac{1 + K_a a - \sqrt{(1 + K_a(R_{Tet} + R_{Tup} - a))^2 + 4K_a a}}{2K_a(R_{Tet} + R_{Tup})}. \tag{19}$$

Experimental data showed that TetR and TetR-Tup1 have different repression efficiencies [11], 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. The SBML version of the model is available as Additional file 1.

We used the 'deSolve' package [32] to solve the ODE model. The final time for simulations was set to 25×10^6 minutes. To ensure that the model reached steady state, we increased simulation time when the relative variation in Citrine observed over the last 500 min exceeded 10^{-10}.

WTC design problem

The steady-state dose-response curve as a reference behavior takes the aTc concentration a as a constant input $u(t) \equiv a$, and yields a constant response $D^{\mathrm{ref}}(a) \equiv D^{\mathrm{ref}}(\tau; a)$ for all τ. We encode the high-IDR, high-ODR objective by defining $(a, D^{\mathrm{ref}}(a))$ to be the straight line between (0 nM, 0 nM) and (150 nM, 60 nM).

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 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 (every 25 nM) between 0 and 150 nM:

$$s(\beta) = \sqrt{\frac{1}{N} \sum_{k=1}^{N} \left(D(\beta, a_k) - D^{\mathrm{ref}}(a_k) \right)^2}. \tag{20}$$

Sampling parameter spaces for the WTC

We defined the pseudo-likelihood for the individual parameter space as:

$$l(\beta) = \mathbb{1}(s(\beta) \leq \varepsilon) \tag{21}$$

with $\varepsilon \in \{6\,\mathrm{nM}, 3\,\mathrm{nM}\}$, therefore sampling uniformly the viable region $V^{\mathrm{ind}} = \{\beta \in B \mid s(\beta) \leq \varepsilon\}$.

For the naive sampling approach, the pseudo-likelihood for the population parameter space was:

$$L(\gamma) = \mathbb{1}(c(\gamma) \leq \delta) \tag{22}$$

with $\delta = 0.2$. We then obtain uniformly distributed samples from the population viable space $V^{\mathrm{pop}} = \{\gamma \in \Gamma \mid c(\gamma) \leq \delta\}$. Note that, as $c(\gamma)$ depends on the value of ε (Eq. 5), $L(\gamma)$ and the associated population viable space will also depend on its value.

To compute the population pseudo-likelihood (Eq. 22), however, we need to approximate $c(\gamma)$, as it is the functional of a distribution (here, 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 $\varepsilon \in \{6\,\mathrm{nM}, 3\,\mathrm{nM}\}$.

Note that we are interested in the resulting distribution of the individual costs and not in describing P_γ. Thus, although we consider six cell-to-cell variable parameters, a sample size of $N = 300$ individual parameters 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. 2C, 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 $N = 300$.

For SLMCMC sampling, the pseudo-likelihood for the population parameter space was:

$$L(\gamma) = P_\gamma^\kappa(s(\beta) \leq \varepsilon) \tag{23}$$

with $\kappa > 1$ and, again, $\varepsilon \in \{6\,\mathrm{nM}, 3\,\mathrm{nM}\}$. This form of the population pseudo-likelihood will result in a higher density of samples in regions of space that have a lower population cost. This is different from the behavior of the uniform population pseudo-likelihood used for naive sampling, which does not discriminate between two populations as long as both their population costs are below the threshold δ.

For the scalar covariance matrix, we ran parallel tempering with inverse temperatures $\{3\}, \{3, 5\}, \{3, 7\}, \{3, 9\}$, therefore requiring a (maximum) total of respectively 3, 10, 14 and 18 individual samples per step, accounting for the swapping step. In the last three cases,

the inverse temperature 3 was present to improve mixing only, and we kept only the samples associated with the higher inverse temperature. We sampled 200,000 populations, removed the first half as burn-in, and then took 600 populations by regularly thinning the samples. An approximation of $c(\gamma)$ was computed using the same approach as for naive sampling. Non-viable populations (i.e., with $c(\gamma) > 0.2$) were discarded, and we compared the fractions of rejected populations for the different κ values. For the diagonal covariance matrix problem, we fixed $\kappa = 7$ because this choice yielded most samples with population cost distributed between 0 and δ, and a small fraction (always smaller than 6.5%) higher than δ. This was assessed *a posteriori* on obtained samples as well.

Sampling efficiency

To compute the Effective Sample Size (ESS) for each dimension, we used the R package 'coda' [33]. The numbers shown in Table 2 were computed only once on a standard laptop, and are subject to variation. Additionally, we expect the numbers involving ESS to be highly problem-dependent, even on the same machine using the same software.

Abbreviations

ODE	Ordinary differential equation
CSL	Continuous stochastic logic
NLME	Non-linear mixed-effects
p.d.f.	Probability density function
MCMC	Markov chain Monte-Carlo
SLMCMC	Stochastic likelihood Markov chain Monte-Carlo
WTC	Well-tempered controller
aTc	AnhydroTetracycline
IDR	Input dynamic range
ODR	Output dynamic range
CV	Coefficient of variation
ESS	Effective sample size
MinESS	Minimal effective sample size
PT	Parallel tempering

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05538-z.

Additional file 1. SBML model for the WTC This SBML file describes the model used to estimate the WTC's parameters. The C file used to simulate the model during sampling is provided in the Gitlab repository https://gitlab.com/csb.ethz/slmcmc_applications.

Additional file 2: Figure S1. Viable samples in the population parameter space, scalar covariance matrix, obtained with the naive method. Symbols as Fig. 4A.

Additional file 3: Figure S2. Viable samples in the population parameter space, diagonal covariance matrix. Samples in all two-dimensional projections of the parameter space, obtained with the SLMCMC algorithm. Orange dots: viable samples for the threshold on the individual cost $\varepsilon = 6\,\mathrm{nM}$; red dots: viable samples for $\varepsilon = 3\,\mathrm{nM}$; black dots: populations with cost $c(\gamma) > 0.2$, representing less than 4.5% of samples. All parameters are in $\log_{10}$-scale.

Author contributions
BT, HMK and JS conceived the population design framework and designed the WTC case-study. BT designed and implemented the SLMCMC algorithm. EYB designed the WTC model and estimated the parameters from experimental data. All authors wrote the manuscript. All authors read and approved the final manuscript.

Turpin *et al. BMC Bioinformatics* (2023) 24:460

Page 25 of 26

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

References

1. Voigt CA. Synthetic biology 2020–2030: six commercially-available products that are changing our world. Nat Commun. 2020;11:6379. https://doi.org/10.1038/s41467-020-20122-2.
2. Nielsen AAK, Der BS, Shin J, Vaidyanathan P, Paralanov V, Strychalski EA, Ross D, Densmore D, Voigt CA. Genetic circuit design automation. Science. 2016. https://doi.org/10.1126/science.aac7341.
3. Barnes CP, Silk D, Sheng X, Stumpf MPH. Bayesian design of synthetic biological systems. Proc Natl Acad Sci. 2011;108(37):15190–5. https://doi.org/10.1073/pnas.1017972108.
4. Lormeau C, Rudolf F, Stelling J. A rationally engineered decoder of transient intracellular signals. Nat Commun. 2021;12(1):1886. https://doi.org/10.1038/s41467-021-22190-4.
5. Ryan EG, Drovandi CC, McGree JM, Pettitt AN. A review of modern computational algorithms for Bayesian optimal design. Int Stat Rev. 2016;84(1):128–54. https://doi.org/10.1111/insr.12107.
6. Karamasioti E, Lormeau C, Stelling J. Computational design of biological circuits: putting parts into context. Mol Syst Des Eng. 2017;2(4):410–21.
7. 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. 2019;8(1):15–2611. https://doi.org/10.1016/j.cels.2018.12.007.
8. Češka M, Dannenberg F, Paoletti N, Kwiatkowska M, Brim L. Precise parameter synthesis for stochastic biochemical systems. Acta Informatica. 2017;54(6):589–623. https://doi.org/10.1007/s00236-016-0265-2.
9. Lavielle M. Mixed effects models for the population approach: models, tasks, methods, and tools. CPT Pharmacomet Syst Pharmacol. 2015;4:1. https://doi.org/10.1002/psp4.10.
10. Geyer CJ. Markov chain Monte Carlo maximum likelihood. Interface Foundation of North America (1991). Accepted: 2010-02-24T20:38:06Z. https://www.stat.umn.edu/geyer/f05/8931/c.pdf. Accessed 08 June 2021.
11. Azizoğlu A, Brent R, Rudolf F. A precisely adjustable, variation-suppressed eukaryotic transcriptional controller to enable genetic discovery. bioRxiv 2019-1212874461 (2020). https://doi.org/10.1101/2019.12.12.874461.
12. Lormeau C, Rybiński M, Stelling J. Multi-objective design of synthetic biological circuits. IFAC-PapersOnLine. 2017;50(1):9871–6. https://doi.org/10.1016/j.ifacol.2017.08.1601.
13. Turpin B, Bijman EY, Kaltenbach HM, Stelling J. Population design for synthetic gene circuits. Lecture notes in bioinformatics. Berlin: Springer; 2021. p. 181–97. https://doi.org/10.1007/978-3-030-85633-5_11.
14. Mannan AA, Liu D, Zhang F, Oyarzún DA. Fundamental design principles for transcription-factor-based metabolite biosensors. ACS Synth Biol. 2017;6(10):1851–9. https://doi.org/10.1021/acssynbio.7b00172.
15. Haario H, Saksman E, Tamminen J. An adaptive metropolis algorithm. Bernoulli. 2001;7(2):223–42. https://doi.org/10.2307/3318737.
16. Bijman EY, Kaltenbach H-M, Stelling J. Experimental analysis and modeling of single-cell time-course data. Curr Opin Syst Biol. 2021;28: 100359. https://doi.org/10.1016/j.coisb.2021.100359.
17. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. J Comput Graph Stat. 1998;7(4):434–55. https://doi.org/10.1080/10618600.1998.10474787.
18. Sakurai Y, Hori Y. Optimization-based synthesis of stochastic biocircuits with statistical specifications. J R Soc Interface. 2018;15:10. https://doi.org/10.1098/rsif.2017.0709.
19. Bonny AR, Fonseca JP, Park JE, El-Samad H. Orthogonal control of mean and variability of endogenous genes in a human cell line. Nat Commun. 2021;12(1):1–9. https://doi.org/10.1038/s41467-020-20467-8.
20. Vignoni A, Oyarzún DA, Picó J, Stan GB. Control of protein concentrations in heterogeneous cell populations. In: 2013 European control conference (ECC) (2013). p. 3633–9. https://doi.org/10.23919/ECC.2013.6669828.
21. Drugan MM, Thierens D. Evolutionary Markov chain Monte Carlo. Lecture notes in computer science. Berlin: Springer; 2004. https://doi.org/10.1007/978-3-540-24621-3_6.

22. Rasmussen CE, Williams CKI. Gaussian processes for machine learning. Adaptive computation and machine learning. Cambridge: MIT Press; 2006.
23. 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. 2018;6(5):593-603.e13. https://doi.org/10.1016/j.cels.2018.04.008.
24. Ghosh S, Birrell P, De Angelis D. Variational inference for nonlinear ordinary differential equations. In: Banerjee A, Fukumizu K (editors) Proceedings of the 24th international conference on artificial intelligence and statistics. Proceedings of machine learning research, PMLR; 2021. p. 2719–27. http://proceedings.mlr.press/v130/ghosh21b.html.
25. Martín PV, Munz MA, Pigolotti S. Bet-hedging strategies in expanding populations. PLOS Comput Biol. 2019;15(4):1006529. https://doi.org/10.1371/journal.pcbi.1006529.
26. Müller P, Sansó B, De Iorio M. Optimal Bayesian design by inhomogeneous Markov chain simulation. J Am Stat Assoc. 2004;99(467):788–98. https://doi.org/10.1198/016214504000001123.
27. Marjoram P, Molitor J, Plagnol V, Tavaré S. Markov chain Monte Carlo without likelihoods. Proc Natl Acad Sci. 2003;100(26):15324–8. https://doi.org/10.1073/pnas.0306899100.
28. Chiachio M, Beck JL, Chiachio J, Rus G. Approximate Bayesian computation by subset simulation. SIAM J Sci Comput. 2014;36(3):1339–58. https://doi.org/10.1137/130932831.
29. Roberts GO, Rosenthal JS. Examples of adaptive MCMC. J Comput Graph Stat. 2009;18(2):349–67. https://doi.org/10.1198/jcgs.2009.06134.
30. Vihola M. Robust adaptive Metropolis algorithm with coerced acceptance rate. Stat Comput. 2012;22(5):997–1008. https://doi.org/10.1007/s11222-011-9269-5.
31. Robert CP, Elvira V, Tawn N, Wu C. Accelerating MCMC algorithms. WIREs Comput Stat. 2018;10(5):1435. https://doi.org/10.1002/wics.1435.
32. Soetaert K, Petzoldt T, Setzer RW. Solving differential equations in R: package deSolve. J Stat Softw. 2010;33(9):1–25. https://doi.org/10.18637/jss.v033.i09.
33. Plummer M, Best N, Cowles K, Vines K. Coda: convergence diagnosis and output analysis for MCMC. R News. 2006;6(1):7–11.

Publisher's Note