Machine learning for optimization of multiscale biological circuits

Charlotte Merzbacher,[1] Oisin Mac Aodha,[1,2] and Diego A. Oyarzún[1,2,3,4]
[1)] *School of Informatics, University of Edinburgh, UK*
[2)] *The Alan Turing Institute, London, UK*
[3)] *School of Biological Sciences, University of Edinburgh, UK*
[4)] *Corresponding author: d.oyarzun@ed.ac.uk*

(Dated: 23 February 2023)

Abstract: Recent advances in synthetic biology have enabled the construction of molecular circuits that operate across multiple scales of cellular organization, such as gene regulation, signalling pathways and cellular metabolism. Computational optimization can effectively aid the design process, but current methods are generally unsuited for systems with multiple temporal or concentration scales, as these are slow to simulate due to their numerical stiffness. We present a machine learning method for the efficient optimization of biological circuits across scales. The method relies on Bayesian optimization and nonparametric statistical models to learn the shape of a performance landscape and iteratively navigate the design space towards an optimal circuit. This strategy allows the joint optimization of both circuit architecture and parameters, and hence provides a feasible approach to solve a highly non-convex optimization problem in a mixed-integer input space. We illustrate the applicability of the method on gene circuits designed to control biosynthetic pathways with multiple interacting scales and various performance objectives. We tested the method on four model pathways of increasing complexity, which highlights its ability to solve large multiscale problems with more than 20 species and circuit architectures. The method can serve as an efficient *in silico* screening method for circuit architectures prior to experimental implementation.

I. INTRODUCTION

The design of molecular circuits with prescribed functions is a core task in synthetic biology[1]. These circuits can include components that operate across various scales of cellular organization, such as gene expression, signalling pathways[2] or metabolic processes[3]. Computational methods are widely employed to discover circuits with specific dynamics[4–6] and, in particular, optimization-based strategies can be employed to search over design space and single out circuits predicted to fulfil a desired function[7–10]. However, circuit design requires the specification of circuit architecture, i.e. the circuit "wiring diagram", as well as the strength of interactions among molecular components. Since circuit architectures are discrete choices and molecular interactions depend on continuous parameters such as binding rate constants, circuit design leads to mixed-integer optimization problems that can be notoriously difficult to solve[11]. Moreover, when circuits operate across multiple scales, their computational models become numerically stiff[12], resulting in extremely slow simulations that make their mixed-integer optimization challenging or even impossible to solve.

Previous work on computational circuit design has largely focused on genetic circuits that operate in isolation from other layers of the cellular machinery (Figure 1A). A range of techniques have been employed to identify functional circuits, including exhaustive search[4–6,13], computational optimization[7,8], Bayesian design[14,15], and machine learning[9,16]. While these methods differ on their specific modelling strategies and assumptions, they all require computational simulations at many, typically thousands to millions, parameter value locations in the design space. In the case of multiscale circuits, the computational cost of simulations grows sharply and limits the application of current optimization. As a result, often these multiscale systems cannot be simulated at the large number of locations in the design space needed by computational search algorithms.

A notable example of this challenge appears in genetic circuits for dynamic control of metabolic pathways[17–21]. These systems are receiving substantial attention thanks to several successful implementations that improved yields as compared to classic techniques in metabolic engineering[22–24]. The key principle is to put enzymatic genes under the control of metabolite-responsive mechanisms that couple heterologous expression to the concentration of a pathway intermediate[3]. This creates feedback loops between enzyme expression and pathway intermediates that allow controlling pathway activity in response to upstream changes in growth conditions or precursor availability. Such dual genetic-metabolic systems are particularly challenging to simulate efficiently, as metabolites and enzymes vary in different timescales, from milliseconds (enzyme kinetics) to minutes (enzyme expression), and they also appear in vastly different concentrations; in bacteria enzymes are typically expressed in nanomolar concentrations, whilst metabolites are found typically above the millimolar range[25]. Moreover, the implementation of these systems is costly and requires substantial experimental fine-tuning. As a result, a central task prior to implementation is the choice of a suitable feedback control loops between metabolites and enzymatic genes, and the strength of interactions between metabolites and actuators of gene expression such as transcription factors[26] or riboregulators[27]. The design of control architectures is particularly important, because there are many ways of building similar control

loops[28], for example by employing combinations of transcriptional activators and repressors[29,30], that may differ in their performance and cost of implementation.

Here, we present a fast and scalable machine learning approach for optimization of multiscale circuit architectures and parameters (Figure 1A). The method is based on Bayesian optimization coupled with differential equation models, and we highlight its utility in various models of metabolic pathways under genetic feedback control[31]. Using a toy example for a simple pathway, we first show that the method converges rapidly and outperforms other optimizers by a substantial margin. We then consider real-world models of metabolic pathways in *Escherichia coli* for the production of several relevant precursors: glucaric acid[32], fatty acids[29], and p-aminostyrene[30]. We use these pathways to illustrate how the speed of our method enables screening optimal designs in realistic design tasks that would otherwise be infeasible to compute, including the impact of uncertain enzyme kinetic parameters, the use of layered architectures that combine metabolic and genetic control, and the optimization of a complex model with 23 differential equations, 27 candidate control architectures, and 19 parameters to be optimized. Our parameterization of the discrete architecture space enables both large numbers of possible architectures and removal of oscillatory or positive feedback architectures from consideration. Machine learning methods such as this one can speed the construction of synthetic biological circuits and present a novel approach to design space exploration.

II. RESULTS

A. Bayesian optimization for joint optimization of circuit architecture and parameters

In general, a circuit design task can be stated as the following mixed-integer optimization problem:

$$\begin{aligned} &\min_{p_d,\, p_c} J(x, p_c, p_d), \\ &\text{subject to:} \\ &\qquad \mathrm{d}x/\mathrm{d}t = h(x), \\ &\qquad p_c \in \mathcal{C},\ p_d \in \mathcal{D}, \end{aligned} \tag{1}$$

where $J(x, p_c, p_d)$ is a performance objective to be optimized over a space of continuous parameters p_c and a discrete set of circuit architectures p_d. The function $h(x)$ is a nonlinear function describing the dynamics of x. The ODE in (1) describes the temporal dynamics of circuit components and are typically built from mass balance equations. Common examples of continuous parameters in applications are binding affinities between DNA and regulatory proteins, or the strength of protein-protein interactions. Conversely, circuit architecture would typically involve various combinations of positive and negative feedback loops among molecular species. We have stated the problem as minimization of J, but similar formulations can be posed as a maximization problem.

To illustrate the utility of the method in a range of design problems, we focus on genetic control circuits for metabolic pathways that synthesize high-value products. In this case, the ODE in (1) contains two sets of equations:

$$\begin{aligned} \mathrm{d}s/\mathrm{d}t &= f(s, e) - \lambda s, \\ \mathrm{d}e/\mathrm{d}t &= u(s, p_c, p_d) - \lambda e, \end{aligned} \tag{2}$$

where s and e are vectors of metabolite and enzyme concentrations, respectively. Both sets of species change in vastly different timescales; metabolic reactions operate in the millisecond range or faster[37], whilst enzyme expression changes in the scale of minutes or longer. Moreover, metabolites and enzymes are also present in different ranges of concentrations, from nM for enzymes to mM and higher for metabolites[25]. As a result, simulation of the ODE in (2) is time consuming, particularly when this needs to be done many times as part of an optimization-based search. The term $f(s, e)$ describes the mass balance relations between pathway intermediates, while the parameter λ models the dilution effect by cell growth. The vector $u(s, p_c, p_d)$ describes the enzyme expression rates controlled by some pathway intermediates, and typically take the form of sigmoidal dose-response curves that lump together processes such as metabolite-TF or metabolite-riboregulator interactions[26]. The continuous parameters p_c model the dose-response curves of the feedback mechanisms, whereas the discrete parameters p_d specify the control architecture.

The performance objective J can be flexibly used to model common design goals such as production flux, yield or titer, as well as cost-benefit tasks that balance production with the deleterious impact of the pathway on the physiology of the host. To first establish a baseline for the performance of our method, we employed a simple toy pathway model that displays common features found in real metabolic pathway (Figure 1C). The model includes a metabolic branch point through a heterologous pathway with two enzymatic steps. As a performance objective we considered the minimization of

$$J = \alpha_1 J_{\text{prod}} + \alpha_2 J_{\text{cost}}, \tag{3}$$

where J_{prod} was designed so that its minimization is equivalent to maximization of the production flux, and J_{cost} penalizes total amount of enzyme expressed during the culture. The parameters α_1 and α_2 are positive weights used to control the balance between the costs and benefits of expressing the heterologous pathway. Details on the objective function can be found in the Supplementary Information.

We considered the four control architectures shown in Figure 1C, which include open loop control as well as three different implementations of negative feedback control using a metabolite-responsive transcription factor. Negative feedback is widely employed in gene cir-

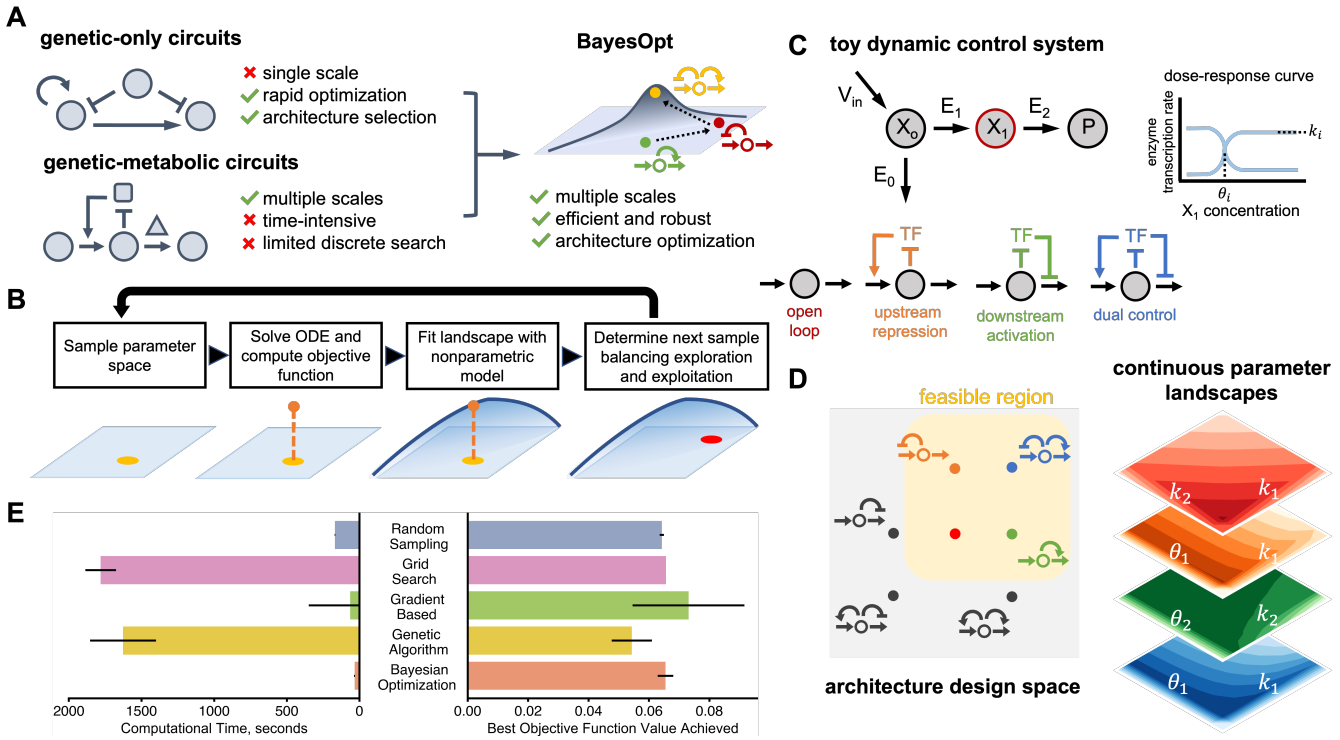

FIG. 1. **Bayesian optimization for the design of circuit architectures and parameters.** **(A)** Previous optimization methods have focused on genetic circuits in isolation from other cellular processes. For multiscale circuits, optimization approaches become infeasible due to the difficulty of simulating stiff dynamical systems in many locations of the design space; a common example of such multiscale systems are gene circuits that control metabolic production[3]. We propose the use of Bayesian optimization (BayesOpt) for efficient optimization of architectures and parameters in multiscale circuits. **(B)** Schematic of a mixed-integer Bayesian optimization loop; the objective function is regarded a random variable to be optimized over an input space comprised of continuous parameters and a set of discrete circuit architectures. At each iteration, the algorithm computes the value of the objective function from the solution of an ordinary differential equation (ODE) model at a single location in the input space. The algorithm learns the shape of the objective landscape using a nonparametric statistical model[33], which is employed to propose a new location in the input space through an "acquisition function" designed to balance exploration and exploitation of the input space; more details in Methods. The algorithm iteratively learns the shape of the performance landscape until convergence to a global optimum. **(C)** Example metabolic pathway under gene regulation. We consider three negative feedback architectures plus open loop control. The intermediate X_1 binds a transcription factor (TF) that controls the expression of pathway enzymes, either as an activator or repressor. The TF dose-response curve (at right) is described by three parameters, k_i, θ_i, and n, where $i = 1, 2$.. The aim is to find designs with optimal architecture and dose-response parameters (k_i, θ_i); for simplicity the Hill coefficient was fixed to $n_i = 2$. **(D)** Performance landscapes of the four feasible circuit architectures. We exclude architectures with positive feedback loops as these are prone to multistability[34]. The shape of the performance landscape defined in (3) shows substantial variation across the four architectures. This leads to a highly non-convex mixed-integer optimization problem. Heatmaps show the value of the objective J computed on a regular grid of the indicated parameters. **(E)** Comparison of BayesOpt against other strategies using the toy model as a benchmark; lower objective function values are better. Shown are the results for random sampling ($N = 1,000$ samples), grid search ($N = 40,000$), a genetic algorithm[35] ($N = 100$ individuals, $N = 1000$ generations), and a gradient-based optimizer to find optimal continuous parameter values for each architecture[36].

cuits as it has substantial benefits in terms of robustness and performance, and their properties have been extensively studied in the literature[38-40]. To illustrate the challenge of jointly optimizing circuit architecture and parameters, in Figure 1D we show a schematic of the design space. The four control architectures under consideration reside at different discrete points in the architecture space. Within each architecture the continuous performance landscape computed as a function J of the dose-response parameters p_c shows substantial variations. We observe cases with convex landscapes with a clear optimum (e.g. dual control) and landscapes with flat basins where most optimization algorithms would struggle to find the optimum (e.g. downstream activation). When searching over the space of architectures and parameters simultaneously, the problem becomes a mixed-integer, non-convex optimization that is extremely challenging to solve with traditional approaches.

We implemented a BayesOpt routine to jointly compute the architecture (p_d) and dose-response parameters (p_c) that minimize the performance objective in Eq. (3). We benchmarked its performance against several other methods, including a random search, an exhaustive grid search, a gradient based method, and a genetic algorithm (Figure 1E). The algorithm was able to compute optimal solutions rapidly (average 27 seconds per run across 100

runs) and robustly (standard deviation less than 2.5% of the mean optimal objective function value). BayesOpt runs significantly faster than the other methods, and provides over a 30-fold improvement over a genetic algorithm. The accuracy of the optimum, quantified by the minimal value of the objective function, is on average 11.4% worse than the genetic algorithm, but this falls within the variation of the latter across several runs. We also note that the traditional gradient-based optimizer proved unreliable and failed to converge on 14.5% of runs.

One key advantage of Bayesian methods is that they are not gradient-based, and therefore are not constrained to navigate the space smoothly in the direction of steepest descent. Gradient-based methods can get trapped in local minima and struggle to find the global optimum, especially in highly nonconvex landscapes like the ones presented here. In contrast, BayesOpt does not converge by chasing minima directly but rather by modelling the entire objective function landscape, which results in rapid and reliable results. The method can perform multiple "jumps" between distant locations in the discrete-continuous search space, where each subsequent sample is selected to maximize the expected improvement on the best sample found so far.

The speed of our approach enables the computation of large solution ensembles under model perturbations such as sweeps of key model parameters. In addition, our method can search high-dimensional mixed-integer design spaces. We next illustrate the versatility of the approach in a range of relevant real-world pathways that require solving the optimization problem for large samples of parameter values.

B. Robustness of control circuits to uncertainty in enzyme kinetic parameters

A challenge in building pathway models is the substantial uncertainty on the enzyme kinetic parameters; this is particularly critical for pathways that include regulatory mechanisms such allostery or product inhibition, which are often poorly characterized. Databases such as BRENDA[42] often have insufficient data on enzyme kinetics for a particular host strain or substrate of interest. Since pathway dynamics can strongly depend on enzyme kinetics, the parametric uncertainty requires extensive sweeps of kinetic parameters to determine the robustness of a specific control architecture deemed to be optimal.

We focused on a pathway for synthesis of glucaric acid in *E. coli* (Figure 2A), a key precursor for many downstream products[32]. The pathway branches from glucose-6-phosphate (g6p) in upper glycolysis and contains three enzymatic steps (Ino1, SuhB, and MIOX). Doong and colleagues implemented a dynamic control circuit using the dual transcriptional regulator IpsA which responds to the intermediate myo-inositol (MI)[21]. The pathway enzyme MIOX is allosterically activated by its own precursor, and one intermediate, MI, can be exported to the extracellular space. We employed a previously developed ODE model[10] that was parameterized using a combination of enzyme kinetic data and omics measurements, and considered the same four control architectures as in the previous example, including various alternative implementations of negative feedback control.

The results in Figure 2B show a typical run of the optimizer when using the cost-benefit objective in Eq. (3) (details in Supplementary Information), together with the fraction of samples in which the algorithm explored each control architecture across the successive iterations. The optimal architecture (dual control in this case) was found quickly and the algorithm was able to further decrease the value of the objective function by exploring the space of dose-response parameters of IpsA. We observe that as the iterations progress, the algorithm shows a remarkable ability to explore other architectures despite their larger objective function values, thus highlighting the global nature of the algorithm.

To explore the impact of uncertain enzyme kinetics, we perturbed the parameters of the rate-limiting MIOX allosteric reaction:

$$\begin{aligned} V_{\mathrm{MIOX}} &= \frac{V_{\mathrm{m,\,eff}}\mathrm{MI}}{k_{\mathrm{m,\,MIOX}} + \mathrm{MI}}, \\ \text{given } V_{\mathrm{m,\,eff}} &= V_{\mathrm{m,\,MIOX}} \frac{1 + a_{\mathrm{MIOX}}\mathrm{MI}}{k_{\mathrm{a,\,MIOX}} + \mathrm{MI}}, \end{aligned} \tag{4}$$

where $V_{\mathrm{m,\,MIOX}}$ is the maximum rate of reaction, $k_{\mathrm{m,\,MIOX}}$ is the Michaelis-Menten constant, and $k_{\mathrm{a,\,MIOX}}$ and a_{MIOX} are allosteric activation constants. We solved the optimization problem for 1000 combinations of these three parameters, which took under 16 hours in a standard laptop machine. Perturbing the kinetic parameters of the glucaric acid pathway did not significantly affect the minimum objective function value achieved, indicating that the optimum is robust to uncertainty in the kinetic parameters (Figure 2C). However, the mean optimal objective function value was not significantly higher among the perturbed samples. We found that the dual control architecture was chosen as optimal in more than 85% of samples (Figure 2D). We thus sought to examine the optimal dose-response parameters of this architecture in more detail.

The maximal enzyme expression rates (k) and regulatory thresholds (θ) control the shape of the dose-response curves. As shown in Figure 2E, we found that the upstream repressive loop and downstream activatory loop had different optimal dose-response curves, corresponding to different optimal values of the continuous parameters. Optimal values of the upstream repression threshold θ_1 are low (mean value 0.64) and compressed into a narrow range compared to the optimal values of downstream repression threshold θ_2 (mean value 7.24). The larger standard deviation on θ_2 values is reflected in the wider confidence interval over the dose response curve for the downstream loop as compared to the upstream loop. Experimental fine-tuning of a dual control circuit might

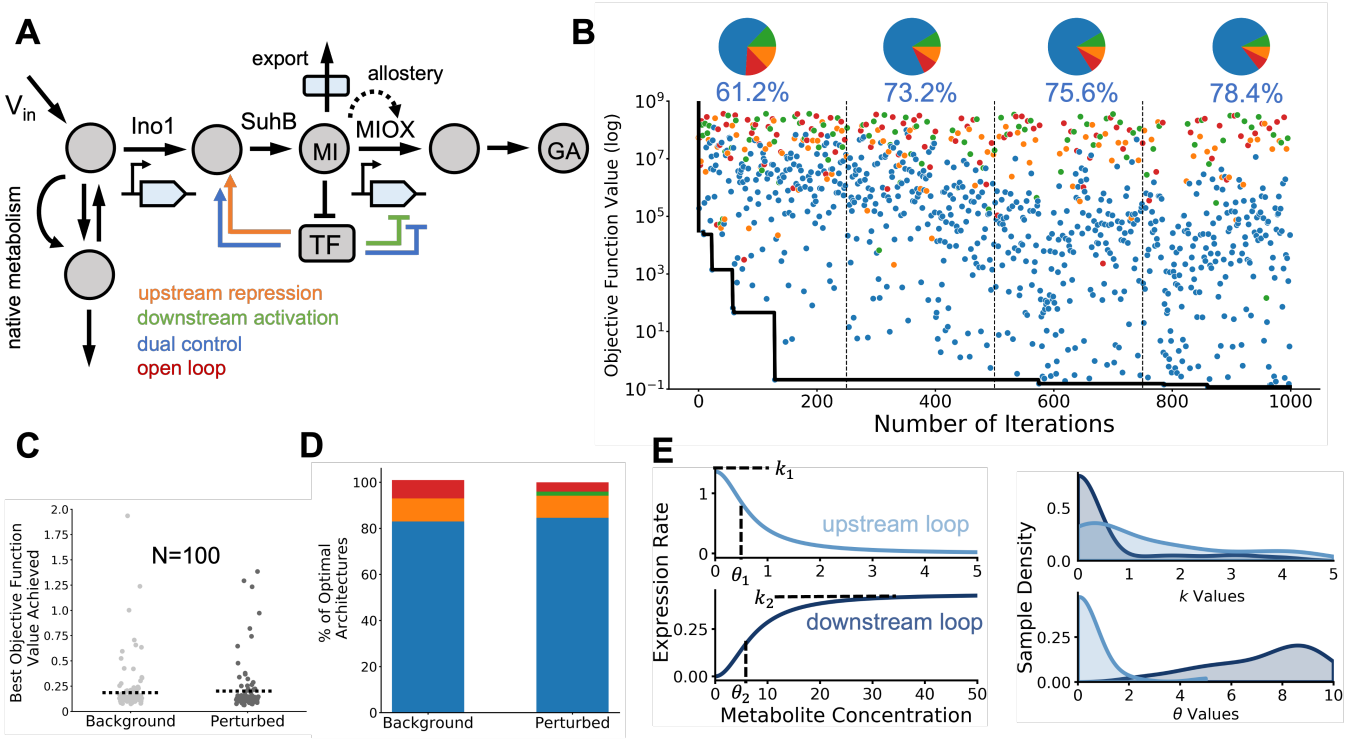

FIG. 2. **Robustness optimal circuits to parameter uncertainty. (A)** Schematic of a dynamic pathway for production of glucaric acid in *Escherichia coli*[21]. The pathway includes allosteric inhibition and export of an intermediate to the extracellular space. The core pathway components myoinositol (MI) and glucaric acid (GA) are modelled explicitly, as are the enzymes Ino1 and MIOX. The enzyme SuhB is not rate-limiting and is not modeled explicitly. **(B)** Sample run of the BayesOpt algorithm for 1,000 iterations of the loop in Figure 1B. Black line shows the descent on the value of the objective function. Dots show all samples colored by architecture; pie charts show the fraction of architectures explored by the algorithm, and the fraction of samples taken from the majority architecture (dual control). The first quarter of the run had the most exploration of architectures other than dual control, with 38.6% of samples coming from non-majority architectures. This percentage steadily decreased over the iterations but did not drop below 20%, illustrating the global nature of the optimization routine. **(C)** To examine the robustness of the optimal solutions to parameter uncertainty, we computed optimal solutions for many perturbed parameters of the allosteric activation of MIOX by its substrate myo-inositol (MI). Strip plot show the best objective function values achieved for background and perturbed kinetic parameters ($V_{\mathrm{m,MIOX}}, a_{\mathrm{MIOX}}, k_{\mathrm{a,MIOX}}$) in Eq. (4). Kinetic parameters were perturbed using Latin Hypercube sampling[41] on the range (-100%, +100%) of the nominal values (Supplementary Information). We observed little difference between between background and perturbed values; dashed line denotes the mean value of the objective function. Only one of the $N = 100$ runs for perturbed parameters failed to converge the optimum. **(D)** Optimal architectures across runs with background and perturbed parameter values. Both background and perturbed systems resulted in over 80% of runs selecting dual control as the optimal architecture. **(E)** Average dose-response curves and distribution of optimal parameters for the dual control architecture with perturbed allosteric parameters. The repressive and activatory loops have substantially different mean dose-response curves on average. The distributions of the dose-response parameters (right) show important variations in their mean and dispersion. The parameter k_i and θ_i determine the maximal enzyme expression rate and regulatory threshold, respectively.

target parameters with optimal values with a wide range, such as k_1, as varying these parameters is less likely to impair circuit function. Overall, these results show the robustness of the glucaric acid dual control system to kinetic parameter uncertainty and demonstrate the possibilities enabled by the speed of BayesOpt.

C. Exploration of alternative objective functions

In the previous case studies we employed a cost-benefit objective designed to account the tradeoff between heterologous production and the cost of expressing pathway enzymes, as in Eq. 3. To demonstrate the flexibility of the method with other objective functions, here we consider the optimization of the temporal trajectories of pathway metabolites.

We focussed on the joint optimization of the rise time and overshoot in a model of a fatty acid production pathway considered previously in the literature[29]. Fatty acids are an essential energy source and cellular membrane component. In addition, hydrocarbons derived from fatty acids have attracted attention as a potential biofuel source[19,43]. Recent work engineering metabolic and genetic control loops showed that negative feedback control could speed up the rise to steady-state conditions[29]. The pathway built in literature expressed a thioesterase under transcriptional control, shown as the negative metabolic loop (NML) architecture in Figure 3A. In addition to transcription-factor mediated negative feedback loops,

this model also includes individually implemented direct genetic loops where a repressor is expressed on the same promoter as the enzyme. These two different scales of loops interface with different levels of cellular organization. We explore several control architectures previously proposed in the literature[29] (Figure 3A).

We first considered a similar objective function as in Eq. (3) so as to compare convergence against the previous case studies. We implemented a modified production flux objective which takes the reciprocal of the product flux to convert the optimization to a minimization problem. The pathway cost J_{cost} is measured by summing the expression of all heterologous enzymes and varies across the different architectures. Details on the objective function can be found in the Supplementary Information. A representative optimization run for this objective (Figure 3B) shows that the negative gene loop (NGL, green) and negative metabolic loop (NML, orange) architectures perform, on average, better than the other two architectures. BayesOpt samples taken from the open-loop architecture were, on average, two orders of magnitude worse than samples taken from NML and NGL architectures. Despite such hierarchy of loss values across the four architectures, the method effectively explores all architectures throughout the optimization run.

We next considered the optimization of percent overshoot and rise-time presented in the literature[29]. The percent overshoot objective, J_{os}, measures the maximal deviation of product from its steady state concentration and is defined as the percent difference between the maximum fatty acid concentration and the steady-state fatty acid concentration. The rise-time, J_{rt}, is a measure of how fast fatty acid production rises to steady-state and is defined as the first time point where the fatty acid concentration reaches 50% of the steady-state value, normalized by the total integration time. We minimized the sum of the overshoot and rise time with a scaling weight α:

$$J = \alpha J_{\text{rt}} + J_{\text{os}}. \tag{5}$$

Adjusting α balances the relative importance of the two optimization criteria; details on the calculation of rise time and overshoot are in the Supplementary Information. Higher values of α correspond to optimal circuits with low rise-times, while lower values of alpha prioritize circuits with low overshoot. Rise-time is a measure of circuit speed, while overshoot is a measure of circuit accuracy. We found that when α is varied across several orders of magnitude, the optimal circuits form a optimal tradeoff curve (Figure 3C). Different architectures occupy different parts of the optimal tradeoff curve and display markedly different dynamics. The NML optima occupies a single point in the loss space, indicating that multiple continuous parameter values give the same loss function value for multiple values of alpha. The NML also has the lowest absolute loss function value of all the architectures considered. The NGL and layered negative metabolic loop (LNML) architectures occupy larger ranges on the curve, with NGL giving a low overshoot and LNML a low rise time. The optimal NML circuit has no overshoot but the slowest rise-time, while the LNML has a very rapid rise-time but overshoots the steady-state value by more than double. These opposing tradeoffs demonstrate the importance of balancing multiple circuit design objectives.

D. Scalability to large pathway models

Our previous case studies have been limited to circuits with a single metabolite controlling gene expression and a relatively small number of control architectures. We now study a large model for the synthesis of p-aminostyrene (p-AS), an industrially relevant vinyl aromatic monomer, in *E. coli* (Figure 4A)[44] using a cost-benefit objective similar to Eq. (3) tailored to the specific pathway (details in Supplementary Information). This model has two possible metabolites that can regulate gene expression, namely p-amminocinnamic acid (p-ACA) and p-aminophenylalanine (p-AF), both of which can act as ligands for aptazyme-regulated expression device (aRED) transcription factors[45], and three genes to be controlled. The aRED transcription factors can also act as dual regulators (activators or repressors) on any of the three promoters involved in the pathway. For simplicity, we limit the design space to control architectures without positive feedback loops, as these are prone to bistability[34]. This results in 27 possible control architectures and 19 continuous parameters to be optimized. The model also has a number of additional complexities. It contains operon-based gene expression commonly found in bacterial systems (genes papA, papB, and papC are expressed on the papABC operon), it includes a detailed description of mRNA dynamics and protein folding, which results in a large model with 23 differential equations, and it can also display oscillatory dynamics.

In addition to expression of heterologous enzymes, the accumulation of toxic intermediates is another major source of genetic burden to host organisms. The p-AS model has several sources of toxicity present in the pathway[30,44]. The intermediate p-ACA and the efflux pump used to remove p-ACA from cells are both cytotoxic, while another intermediate, p-AF, leaks from cells. The pathway enzyme L-Amino Acid Oxidase (LAAO) depletes key aromatic amino acid metabolites and creates toxic hydrogen peroxide as a byproduct. The model incorporates these various types of toxicity in the form of a toxicity factor tau. This toxicity factor is of the form

$$\tau = \frac{k_i}{k_i + \frac{\text{pACA}}{t_a} + \frac{\text{P}_{\text{efflux}}}{t_p} + \frac{\text{LAAO}}{t_l}}, \tag{6}$$

where t_l, t_a, and t_p are chemical-specific toxicity factors. Enzyme-induced toxicity t_l scales the key metabolite depletion rate driven by the enzyme LAAO. Metabolite-

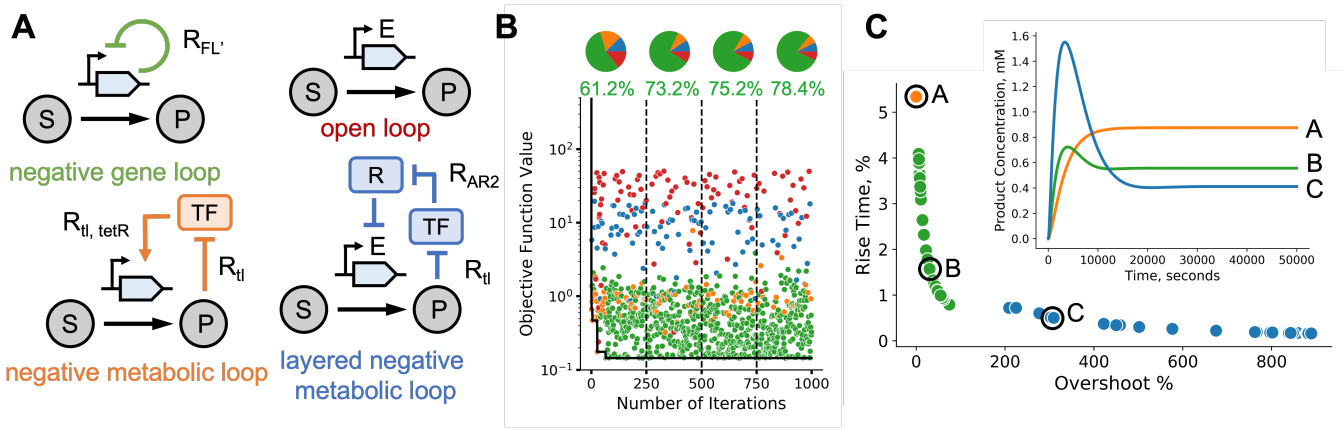

FIG. 3. **Performance landscapes of fatty acid synthesis pathway.** **(A)** Pathway diagrams with various control architectures implemented in *Escherichia coli*[29]. The metabolic loop employs a metabolite-responsive transcription factor, whereas the gene loop includes only a repressor expressed on the same promoter as the enzyme. **(B)** Representative run of BayesOpt with cost-benefit objective showing the best objective function value (black line). All samples are colored by their architecture. Pie charts of each quarter of the run show continued exploration of all architectures despite clear stratification in losses. **(C)** Optimal tradeoff curve between overshoot and rise time. The objective weight α was swept from $\alpha = 0.01$ to $\alpha = 10,000$ and BayesOpt was run for 100 iterations at each α value. The optimal parameter values were used to compute the rise time and overshoot for visualization. The inset shows three sample trajectories illustrating how different optimal architectures navigate the tradeoff between overshot and rise time.

induced toxicity t_a scales the impact of toxic intermediate p-ACA concentration. Finally, protein-induced toxicity t_p reflects the toxicity caused by efflux pump expression. The toxicity factor acts as a scaling coefficient on the pathway synthesis, degradation, and folding reaction rates.

Despite the complexity and size of the p-AS model, we observe that BayesOpt explores many of the 27 possible architectures and converges to a low value of the objective function (Figure 4B); this was also achieved at a reasonable computational cost (mean run time under two minutes). We do not explicitly name the architectures but rather specify them based on the function (e.g. repression, R) and ligand (e.g. p-ACA, 1) at each control point. The best architecture selected in the sample run was R2-R2-A2, but there is no clear best architecture when the optimization is run many times. No architecture is optimal for more than 15% of test runs, demonstrating that there are combinations of architectures and parameter values that achieve similar optimal losses. This broad distribution of optimal architectures appears even if the optimal loss achieved is much smaller than the mean. Several architectures of the model can display oscillatory solutions. We chose to exclude these undesirable solutions from search by applying Scipy peak detection[36] and adding a large regularization term to the objective function of oscillatory solutions.

To investigate the robustness to chemical toxicity, we perturbed the metabolite-induced toxicity t_a and protein-induced toxicity t_p in (6). The optimal loss values were found to be comparable between perturbed and background systems (Figure 4C). Additionally, when projected onto a 2-dimensional space using principal component analysis, the distribution of background parameter values was similar to the distribution of perturbed solutions, indicating that the perturbation did not significantly affect the optimal parameters selected (Figure 4D)[46]. The p-AS pathway lies at the far end of what is currently possible to build experimentally, and thus illustrates the broad applicability of BayesOpt to realistic design tasks in metabolic engineering.

III. DISCUSSION

Synthetic biologists enjoy an unprecedented level of control over biological components. This has allowed the construction of circuits on increased complexity and acting across various levels of biological organization. However, large design spaces and multiple scales of biological organization can become substantial challenges for the rapid design of functional systems.

Gene circuits designed to control metabolic pathways provide an excellent example of such challenges, as they integrate fast metabolic timescales with the much slower dynamics of gene expression. Moreover, the choice of regulators, control points, and control architectures adds multiple degrees of freedom that are infeasible to explore experimentally. Computational methods can aid the design of such systems prior to implementation and serve as tools for *in silico* screening of competing designs that may have similar performance but entail different cost of wetlab implementation.

Previously implemented metabolic control systems have been built primarily based on application-specific knowledge of pathway features[22,23], and there is a lack of computational methods that can accelerate the design cycle. In this paper we presented the application

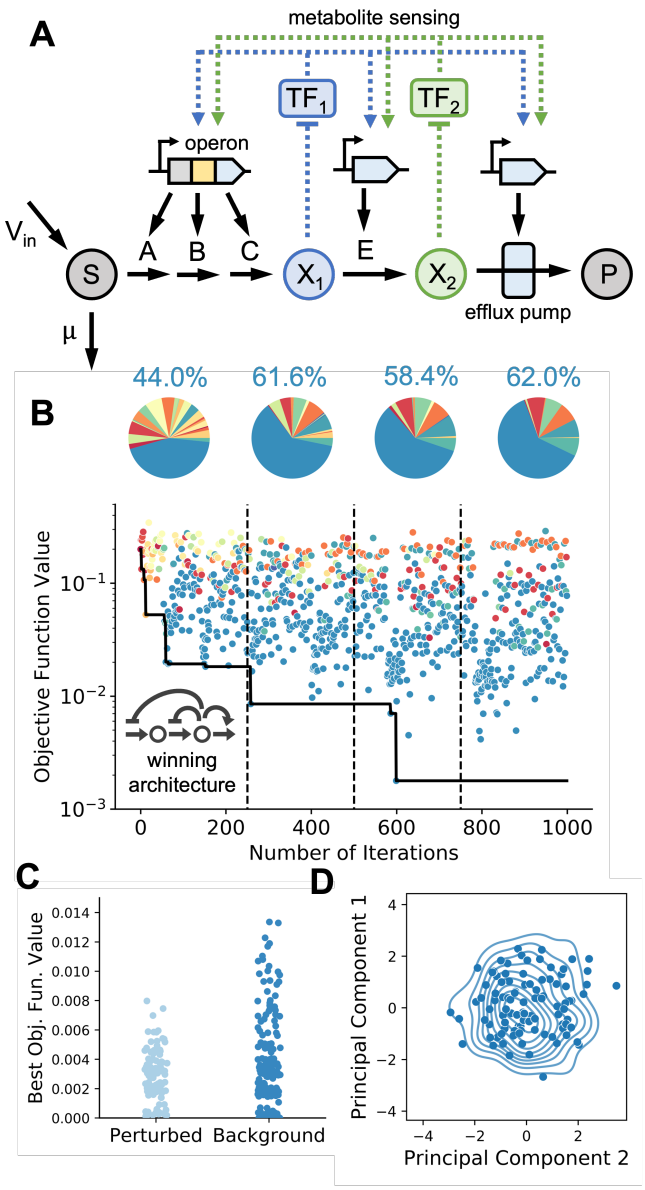

FIG. 4. **Bayesian optimization in a complex pathway. (A)** Schematic of pathway for production of p-aminostyrene[30]. Two intermediates can act as ligands for metabolite-dependent riboregulators, and three promoter sites of control. The optimization problem has 16 continuous decision variables and 27 circuit architectures. The substrate S is converted by enzymes A, B, and C to X_1, which is then converted by E to X_2. The toxic substrate X_2 is then pumped out of the cell via an efflux pump to form the product P. Both X_1 and X_2 can act on the transcription factors TF_1 and TF_2. **(B)** Representative run of the BayesOpt algorithm; the method samples many architectures before settling on the optimal one. Pie charts show continued exploration of a large number of architectures. The winning architecture, R2-R2-A2, is shown as an inset. **(C)** The p-aminostyrene pathway has several forms of substrate, protein, and enzyme toxicity expressed via a toxicity factor τ (see Equation 6). To explore the effects of protein and metabolite toxicity, we perturbed the toxicity factor. Metabolite-induced toxicity was perturbed on the nominal range (10E-3, 10E-4) and protein-induced toxicity on the range (10E-4, 10E1) respectively. Both ranges were selected to match the ranges provided in the literature[30]. Latin Hypercube sampling was used to generate $N = 100$ perturbed parameter values, and the optimal solutions were compared to an equal number of background solutions using the nominal parameter values. **(D)** Visualization of the optimal solutions; scatter plot of principal components of the optimal parameter values for the model with perturbed toxicity parameters (N=100). Contour plots show the background distribution of parameter values.

of a machine learning method for the joint optimization of biological architectures and parameters. We showed the efficiency and scalability of the method in several real-world case studies from metabolic engineering. The p-aminostyrene pathway is more complex than systems typically implemented in literature, which suggests that the method is applicable across a range of real-world design tasks. The method is particularly well suited for cases in which the multiple scales prevent efficient simulation of ODE models in many locations of the design space. We anticipate several novel applications of this work to other problem areas where discovery or tuning of multiscale circuits has been previously infeasible. For instance, this method could be employed to fit temporal circuit dynamics to data or discern which of several discrete circuit mechanisms most closely matches observed behavior.

Machine learning has already proved useful in a range of metabolic engineering tasks[47] and is gaining substantial interest in other areas of synthetic biology[48,49]. In this paper we have shown how such methods can also benefit dynamic pathway engineering, by accelerating the design cycle with model-guided *in silico* screening prior to system prototyping.

IV. METHODS

A. Bayesian optimization

We employed the Bayesian Optimization routine implemented in the Python `HyperOpt` package[33]. Bayesian optimization is commonly employed for hyperparameter tuning in deep neural networks. We employed Expected Improvement as an acquisition function and a tree-structured Parzen estimator (TPE) as a non-parametric statistical model for the loss landscape. We performed a grid search over the TPE hyperparameter γ which controls the balance between exploration and exploitation but found little impact on the algorithm performance; we thus used the default value of $\gamma = 15$ (Supplementary Figure S1).

Constraints on the continuous and discrete decision variables were incorporated directly into the `HyperOpt` search space. At each run of the Bayesian optimization routine, the initial guess for the continuous decision variables were sampled from uniform distributions, with upper and lower bounds were taken from literature[10,30,43]. Architectures were chosen uniformly from the set of architectures without positive feedback loops.

B. Model pathways

We considered four exemplar pathways modelled via ordinary differential equations (ODEs): the toy system in Figure 1C, the glucaric acid pathway in Figure 2A, the fatty acid pathway in Figure 3A, and the p-aminostyrene

pathway in Figure 4A. Table I contains a summary of the four considered models. In all cases, pathway models include ODEs for both metabolites and pathway enzymes. In each case, we define the various control architectures and incorporate them as discrete decision variables in the optimization problem, i.e. p_d in Eq. (1); the continuous decision variables, i.e. p_c in Eq. (1), appear in the expression rates of the pathway enzymes. For the toy model and the glucaric acid pathway, enzyme expression was parameterized using a lumped Hill equation model to describe the interaction between a regulatory metabolite and a transcription factor. For the fatty acid and p-aminostryrene pathways, expression rates were parameterized with bespoke nonlinear functions describing specific biochemical processes. The discrete control architectures were defined in two different ways. For the toy, glucaric acid, and p-aminostyrene models, the architectures were defined using a binary matrix to encode the mode of transcriptional control. For the fatty acid model we instead defined each architecture as a categorical choice and switched between model functions correspondingly. We note that the p-aminostyrene pathway also contains ODEs for mRNA abundance and folded/unfolded proteins. All models and their parameters are described in the Supplementary Information.

The ODE models were solved with `scikit-odes`, a Python wrapper for the sundials suite of solvers[50]. In all cases, the initial concentrations of heterologous pathway enzymes were assumed to be zero. Initial concentrations for native metabolites were determined by first solving a model without the heterologous enzymes up to steady state. Simulation times and initial conditions are detailed in the Supplementary Information for each model.

C. Loss function

In all cases the loss function J in Eq. (3) was instanced to each specific pathway. Generally, the loss is defined as a linear combination of costs and benefits of pathway activity so as to balance opposing design goals commonly found in applications. Since both components of the loss function have different magnitudes, for each model we first swept the weights α_1 and α_2 across many model simulations, and chose values that led to similar values for both components; this prevents the optimizer to bias the search towards low loss values caused by the scaling effects. For the fatty acid model in Figure 3 we also optimized the circuit rise-time and overshoot % defined in Eq. (5). Details on all objective functions can found in the Supplementary Information.

CODE AVAILABILITY

Python code for this paper is available on the Github repository.

ACKNOWLEDGEMENTS

CM and DAO were supported by the United Kingdom Research and Innovation (grant EP/S02431X/1, UKRI Centre for Doctoral Training in Biomedical AI).

REFERENCES

[1] Jennifer a N Brophy and Christopher a Voigt. Principles of genetic circuit design. *Nature Methods*, 11(5):508–520, may 2014.
[2] William M. Shaw, Hitoshi Yamauchi, Jack Mead, Glen Oliver F. Gowers, David J. Bell, David Öling, Niklas Larsson, Mark Wigglesworth, Graham Ladds, and Tom Ellis. Engineering a Model Cell for Rational Tuning of GPCR Signaling. *Cell*, 177(3):782–796.e27, 2019.
[3] Fuzhong Zhang, James M Carothers, and Jay D Keasling. Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids. *Nature Biotechnology*, 30(4):354–9, mar 2012.
[4] Wenzhe Ma, Ala Trusina, Hana El-Samad, Wendell A Lim, and Chao Tang. Defining network topologies that can achieve biochemical adaptation. *Cell*, 138(4):760–773, 2009.
[5] Zhengda Li, Shixuan Liu, and Qiong Yang. Incoherent inputs enhance the robustness of biological oscillators. *Cell systems*, 5(1):72–81, 2017.
[6] Lingxia Qiao, Wei Zhao, Chao Tang, Qing Nie, and Lei Zhang. Network topologies that can achieve dual function of adaptation and noise attenuation. *Cell systems*, 9(3):271–285, 2019.
[7] Madhukar S. Dasika and Costas D. Maranas. OptCircuit: An optimization based method for computational design of genetic circuits. *BMC Systems Biology*, 2:1–19, 2008.
[8] Irene Otero-Muras and Julio R. Banga. Automated Design Framework for Synthetic Biology Exploiting Pareto Optimality. *ACS Synthetic Biology*, 6(7):1180–1193, jul 2017.
[9] Tom W Hiscock. Adapting machine-learning algorithms to design gene circuits. *BMC bioinformatics*, 20(1):1–13, 2019.
[10] Babita K Verma, Ahmad A Mannan, Fuzhong Zhang, and Diego A Oyarzún. Trade-offs in biosensor optimization for dynamic pathway engineering. *ACS Synthetic Biology*, 11(1):228–240, 2021.
[11] Julio Banga. Optimization in computational systems biology. *BMC Systems Biology*, 2(1):47, 2008.
[12] Ernst Hairer and Gerhard Wanner. *Solving Ordinary Differential Equations II: Stiff and differential-algebraic problems*. Springer-Verlag, 1996.
[13] Franco Blanchini, Elisa Franco, and Giulia Giordano. A structural classification of candidate oscillatory and multistationary biochemical systems. *Bulletin of mathematical biology*, 76(10):2542–69, oct 2014.
[14] Mae L Woods, Miriam Leon, Ruben Perez-Carrasco, and Chris P Barnes. A statistical approach reveals designs for the most robust stochastic gene oscillators. *ACS synthetic biology*, 5(6):459–470, 2016.
[15] Javier Gonzalez, Joseph Longworth, David C James, and Neil D Lawrence. Bayesian optimization for synthetic gene design. *arXiv preprint arXiv:1505.01627*, 2015.
[16] Jingxiang Shen, Feng Liu, Yuhai Tu, and Chao Tang. Finding gene network topologies for given biological function with recurrent neural network. *Nature communications*, 12(1):1–10, 2021.
[17] Fuzhong Zhang, James M Carothers, and Jay D Keasling. Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids. *Nature biotechnology*, 30(4):354–359, 2012.
[18] Diego A. Oyarzún and Guy-Bart V Stan. Synthetic gene circuits for metabolic control: design trade-offs and constraints. *Journal of the Royal Society Interface*, 10(78), 2013.

product	parameters (p_c)	architectures (p_d)	metabolites	enzymes	ODEs
toy pathway	4	4	2	2	4
glucaric acid[10,21]	4	4	3	2	5
fatty acid[29]	2	4	1	2	3
p-aminostyrene[30]	19	27	7	6	23

TABLE I. Summary of pathway models studied in this paper. The ODEs in the p-aminostyrene pathway also include mRNA and folding dynamics.

[19] Peng Xu, Lingyun Li, Fuming Zhang, Gregory Stephanopoulos, and Mattheos Koffas. Improving fatty acids production by engineering dynamic pathway regulation and metabolic control. *Proceedings of the National Academy of Sciences*, 111(31):11299–11304, 2014.

[20] Mary J Dunlop, Jay D Keasling, and Aindrila Mukhopadhyay. A model for improving microbial biofuel production using a synthetic feedback loop. *Systems and synthetic biology*, 4:95–104, 2010.

[21] Stephanie J Doong, Apoorv Gupta, and Kristala LJ Prather. Layered dynamic regulation for improving metabolic pathway productivity in escherichia coli. *Proceedings of the National Academy of Sciences*, 115(12):2964–2969, 2018.

[22] Cynthia Ni, Christina V Dinh, and Kristala LJ Prather. Dynamic control of metabolism. *Annual Review of Chemical and Biomolecular Engineering*, 12, 2021.

[23] Di Liu, Ahmad A Mannan, Yichao Han, Diego A Oyarzún, and Fuzhong Zhang. Dynamic metabolic control: towards precision engineering of metabolism. *Journal of Industrial Microbiology and Biotechnology*, 45(7):535–543, 2018.

[24] Christopher J Hartline, Alexander C Schmitz, Yichao Han, and Fuzhong Zhang. Dynamic control in metabolic engineering: Theories, tools, and applications. *Metabolic engineering*, 63:126–140, 2021.

[25] Mona K. Tonn, Philipp Thomas, Mauricio Barahona, and Diego A Oyarzún. Stochastic modelling reveals mechanisms of metabolic heterogeneity. *Communications Biology*, 2(1):108, dec 2019.

[26] Ahmad A. Mannan, Di Liu, Fuzhong Zhang, and Diego A. Oyarzún. Fundamental Design Principles for Transcription-Factor-Based Metabolite Biosensors. *ACS Synthetic Biology*, 6(10):1851–1859, oct 2017.

[27] Li-Bang Zhou and An-Ping Zeng. Exploring Lysine Riboswitch for Metabolic Flux Control and Improvement of L-Lysine Synthesis in *Corynebacterium glutamicum*. *ACS Synthetic Biology*, 4(6):729–734, jun 2015.

[28] Madalena Chaves and Diego A. Oyarzún. Dynamics of complex feedback architectures in metabolic pathways. *Automatica*, 99:323–332, 2019.

[29] Di Liu and Fuzhong Zhang. Metabolic feedback circuits provide rapid control of metabolite dynamics. *ACS synthetic biology*, 7(2):347–356, 2018.

[30] Jason T Stevens and James M Carothers. Designing rna-based genetic control systems for efficient production from engineered metabolic pathways. *ACS synthetic biology*, 4(2):107–115, 2015.

[31] Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[32] Tae Seok Moon, Sang-Hwal Yoon, Amanda M Lanza, Joseph D Roy-Mayhew, and Kristala L Jones Prather. Production of glucaric acid from a synthetic pathway in recombinant escherichia coli. *Applied and environmental microbiology*, 75(3):589–595, 2009.

[33] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR, 2013.

[34] Diego A. Oyarzún and Madalena Chaves. Design of a bistable switch to control cellular uptake. *Journal of The Royal Society Interface*, 12(113):20150618, dec 2015.

[35] Ryan M. Solgi. Geneticalgorithm package. https://pypi.org/project/geneticalgorithm/, 2020.

[36] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[37] Arren Bar-Even, Elad Noor, Yonatan Savir, Wolfram Liebermeister, Dan Davidi, Dan S Tawfik, and Ron Milo. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry*, 50(21):4402–4410, 2011.

[38] Naveen Venayak, Nikolaos Anesiadis, William R Cluett, and Radhakrishnan Mahadevan. Engineering metabolism through dynamic control. *Current opinion in biotechnology*, 34:142–152, 2015.

[39] Christopher Hartline, Ahmad Mannan, Di Liu, Fuzhong Zhang, and Diego Oyarzún. Metabolite sequestration enables rapid recovery from fatty acid depletion in Escherichia coli. *mBio*, 11(2):590943, 2019.

[40] Yuan Zhu, Ying Li, Ya Xu, Jian Zhang, Linlin Ma, Qingsheng Qi, and Qian Wang. Development of bifunctional biosensors for sensing and dynamic control of glycolysis flux in metabolic engineering. *Metabolic Engineering*, 68:142–151, 2021.

[41] Wei-Liem Loh. On latin hypercube sampling. *The annals of statistics*, 24(5):2058–2080, 1996.

[42] I Schomburg, L Jeske, M Ulbrich, S Placzek, A Chang, and D Schomburg. The brenda enzyme information system–from a database to an expert system. *Journal of biotechnology*, 261:194–206, 2017.

[43] Yiming Zhang, Jens Nielsen, and Zihe Liu. Metabolic engineering of saccharomyces cerevisiae for production of fatty acid–derived hydrocarbons. *Biotechnology and bioengineering*, 115(9):2139–2147, 2018.

[44] M Ya Goikhman, NP Yevlampieva, NV Kamanina, IV Podeshvo, IV Gofman, SA Mil'tsov, AP Khurchak, and AV Yakimanskii. New polyamides with main-chain cyanine chromophores. *Polymer Science Series A*, 53(6):457–468, 2011.

[45] Andrew D Ellington and Jack W Szostak. In vitro selection of rna molecules that bind specific ligands. *nature*, 346(6287):818–822, 1990.

[46] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[47] Tijana Radivojević, Zak Costello, Kenneth Workman, and Hector Garcia Martin. A machine learning Automated Recommendation Tool for synthetic biology. *Nature Communications*, 11(1):1–14, 2020.

[48] Pablo Carbonell, Tijana Radivojevic, and Héctor García Martín. Opportunities at the Intersection of Synthetic Biology, Machine Learning, and Automation. *ACS Synthetic Biology*, 8(7):1474–1477, jul 2019.

[49] Evangelos-Marios Nikolados, Arin Wongprommoon, Oisin Mac Aodha, Guillaume Cambray, and Diego A. Oyarzún. Accuracy and data efficiency in deep learning models of protein expression. *Nature Communications*, 13(1):7755, dec 2022.

[50] David J Gardner, Daniel R Reynolds, Carol S Woodward, and Cody J Balos. Enabling new flexibility in the sundials suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software (TOMS)*, 2020.

[51] James M Carothers, Jonathan A Goler, Darmawi Juminaga, and Jay D Keasling. Model-driven engineering of rna devices to quantitatively program gene expression. *Science*, 334(6063):1716–1719, 2011.