# BayesianSSA: a Bayesian statistical model based on structural sensitivity analysis for predicting responses to enzyme perturbations in metabolic networks

Shion Hosoda[1*], Hisashi Iwata[1], Takuya Miura[1], Maiko Tanabe[1], Takashi Okada[2], Atsushi Mochizuki[2], and Miwa Sato[1]

[1]Center for Exploratory Research, Research and Development Group, Hitachi, Ltd., Kokubunji-shi, Tokyo 185-8601, Japan
[2]Laboratory of Mathematical Biology, Institute for Life and Medical Sciences, Kyoto University, Kyoto-shi, Kyoto 606-8507, Japan
[*]To whom correspondence should be addressed.

**Abstract**

**Background:** Chemical bioproduction has attracted attention as a key technology in a decarbonized society. In computational design for chemical bioproduction, it is necessary to predict changes in metabolic fluxes when up-/down-regulating enzymatic reactions, that is, responses of the system to enzyme perturbations. Structural sensitivity analysis (SSA) was previously developed as a method to predict qualitative responses to enzyme perturbations on the basis of the structural information of the reaction network. However, the network structural information can sometimes be insufficient to predict qualitative responses unambiguously, which is a practical issue in bioproduction applications. To address this, in this study, we propose BayesianSSA, a Bayesian statistical model based on SSA. BayesianSSA extracts environmental information from perturbation datasets collected in environments of interest and integrates it into SSA predictions.

**Results:** We applied BayesianSSA to synthetic and real datasets of the central metabolic pathway of *Escherichia coli*. Our result demonstrates that BayesianSSA can successfully integrate environmental information extracted from perturbation data into SSA predictions. In addition, the posterior distribution estimated by BayesianSSA can be associated with the known pathway reported to enhance succinate export flux in previous studies.

**Conclusions:** We believe that BayesianSSA will accelerate the chemical bioproduction process and contribute to advancements in the field.

## 1 Background

Chemical production using microbes, known as chemical bioproduction, has attracted attention as a key technology in a decarbonized society. Chemical bioproduction is expected to be essential for sustainable development [1], such as the production of medicines [2, 3], fuels [4], and foods [5], and the absorption of CO2 [6, 7]. For efficient chemical bioproduction, computational designs of metabolic networks are typically employed to reduce the cost of comprehensive wet lab experiments [8, 9].

One powerful strategy for computational design is to predict changes in metabolic fluxes when up-/down-regulating enzymatic reactions, that is, responses to enzyme perturbations [10, 11]. Up-/down-regulation of enzymatic reactions through genetic manipulations, such as modification, over-

expression, and knockout [12, 13], can alter the metabolic fluxes. Increasing the flux of the target chemical leads to efficient chemical bioproduction, and prediction is an essential step in this process.

There are two types of computational methods for predicting responses to enzyme perturbations. The first is flux balance analysis (FBA)-based methods [14, 15, 16, 17], which maximize an objective function, instead of considering kinetics. When applied to unfamiliar strains, FBA-based methods suffer from dependence on the objective function, which is typically the biomass objective function [18]. Specifically, extensive wet lab experiments need to be conducted to measure phenotypes, such as biomass, of the strain of interest and to identify the objectives of the strain in biological activities [19]. The second is kinetics-based methods, such as sensitivity analysis in kinetic models [20, 21, 22] and structural sensitivity analysis (SSA) [10, 23]. Sensitivity analysis in kinetic models allows predicting quantitative responses to enzyme perturbations. There are global and local sensitivity analysis [24, 22], and the local sensitivity analysis is the more typical method for applying to kinetic models, which is called metabolic control analysis [25] and has been successfully used in many applications [26, 27, 28]. To construct a kinetic model, it is necessary to obtain parameters and functional forms of reaction rates of all reactions in the metabolic network of interest under specific environmental conditions [20, 21]. Therefore, parameter estimation is typically essential for unfamiliar strains, whose known information is rarely available [29]. Even though many parameter estimation methods have been developed [30, 31, 32, 33], it can still be a cumbersome process due to the high dimensionality of the parameter space [30, 34]. In contrast, SSA can predict qualitative responses, which are the signs of responses, to enzyme perturbations only from structural information of the metabolic network. SSA does not need to determine the functional forms and parameters of the reaction rates. In other words, SSA removes the burden of parameter determination entirely. Because of its parameter-free nature, SSA could be widely applicable to chemical bioproduction.

SSA is originally a method to predict qualitative responses of a chemical reaction system to perturbations in enzyme amounts (or activities) only from structural information of the reaction network. In SSA, change in the chemical concentrations/fluxes to a perturbation is given in the form of a rational function of "SSA variables," which are defined as derivatives of reaction rates with respect to chemical concentrations. From their definitions (*cf.* the "Our model" section), the SSA variables vary depending on environmental conditions, such as aerobic/anaerobic states, pH levels, and nutrient availability, as well as individual differences among microbes. By the SSA theory, whether the rational function is zero or non-zero, indicating the absence or presence of a response, is unambiguously determined from the structural information of the network alone. Furthermore, the sign of the rational function, indicating a positive or negative response, could potentially be determined by considering the signs of SSA variables, which are usually deducible from general considerations or biological knowledge. However, if the range of the rational function happens to include zero, its sign becomes undeterminable. For example, when a rational function is a subtraction of two positive variables, its sign depends on the quantitative values of these two variables. We refer to such structurally undeterminable response predictions as "indefinite predictions." Indefinite predictions are true as the general consequences of responses drawn from structural information alone, which SSA aims to obtain rather than specific consequences.

However, these indefinite predictions in SSA can present practical challenges in the application to specific species and culture environments for chemical bioproduction, due to the necessity of precisely detecting reactions to up-/down-regulate. They often arise in complex and intertwined metabolic networks. Indeed, the central metabolic pathway analyzed in this study exhibits many indefinite predictions (Supplementary Table S1). As mentioned above, the SSA variables vary depending on the environmental conditions, fluctuating in accordance with the individual differences among microbes. For specific microbial species and culture environments, constraining the possible values of SSA variables on the basis of environmental information may decrease the number of indefinite predictions. This implies that application methods of SSA using environmental information are practically beneficial for chemical bioproduction.

In this study, we propose BayesianSSA, a Bayesian statistical model that extracts environmental information from perturbation datasets collected in environments of interest and integrates it into

SSA predictions. BayesianSSA considers the variables of SSA as stochastic variables, and they are estimated using the perturbation data. Although BayesianSSA introduces new parameters to estimate, it still requires fewer parameters per reaction than kinetic modeling requires. For example, for a one-substrate reaction, BayesianSSA requires one parameter while kinetic modeling with the Michaelis–Menten equation requires two parameters, $V_{\max}$ and $K_{\mathrm{m}}$. In addition, BayesianSSA does not need to explore the functional forms of the reaction rates. The introduction of stochastic variables in BayesianSSA brings two additional advantages. First, it allows for the consideration of the uncertainty caused by individual differences among microbes. The variables of SSA may depend not only on environmental conditions but also on individual differences among microbes, and considering their uncertainty may contribute to predictive performance. Second, it enables a probabilistic interpretation of indefinite prediction by positivity confidence values. These values are defined as the probability that the predictive response is positive. We report the results of applying BayesianSSA to synthetic and real datasets of the central metabolic pathway of *E. coli*. To validate the practicability of BayesianSSA and assess whether BayesianSSA can integrate environmental information into SSA predictions, we compared predictive performances between BayesianSSA and a base method, which is the same as the BayesianSSA model but utilizes an initial prior distribution without incorporating perturbation datasets, as well as a naive Bayes model on these datasets. Utilizing environmental information may enhance predictions for out-of-sample perturbations, where chemical reactions to be perturbed are not included in the sample used to fit BayesianSSA for a given target chemical. To examine this effect, we evaluated predictions of out-of-sample perturbations by BayesianSSA.

# 2 Theoretical background

## 2.1 Structural sensitivity analysis

### 2.1.1 Algorithm

We explain the SSA algorithm [10, 11] briefly in this section. Consider the following ordinary differential equation (ODE):

$$\frac{dx_m}{dt} = \sum_{j=1}^{J} \nu_{m,j} F_j(k_j, \mathbf{x}),$$

where $x_m$ is the concentration of the $m$-th metabolite, $J$ is the number of different reactions, $\nu_{m,j}$ is the $(m, j)$ element of the stoichiometric matrix $\boldsymbol{\nu}$ that indicates metabolites as rows and reactions as columns, $F_j(\cdot)$ is the $j$-th reaction rate function, which shows the reaction flux, $k_j$ is the reaction rate constant of the $j$-th reaction, $\mathbf{x} = (x_1, \ldots, x_M)^{\mathrm{T}}$ is the vector of metabolite concentrations, and $M$ is the number of different metabolites. Under the situation where the system obeys this ODE and the assumption that $F_j$ monotonically increases with respect to $k_j$, the SSA algorithm can derive the rational function of a response to a perturbation of the $j$-th reaction. Here, the perturbation and response are represented as an operation changing $k_j$ to $k_j + \delta k_j$ [11] and the change in metabolite concentrations/reaction fluxes from the initial steady-state to the eventual steady-state after the perturbation, respectively.

The first step of the SSA algorithm is to make a matrix $\mathbf{R}(\mathbf{r})$. The $(j, m)$ element of $\mathbf{R}(\mathbf{r})$, denoted by $r_{j,m}$, is equal to $\partial F_j / \partial x_m$. We write non-zero elements of $\mathbf{R}(\mathbf{r})$ collectively as $\mathbf{r} \in \mathbb{R}^P$, with $P$ being the total number of non-zero values in $\mathbf{R}(\mathbf{r})$. That is, the matrix $\mathbf{R}(\mathbf{r})$ indicates the dependence of reactions to metabolites. For example, $r_{j,m} > 0$ if the $m$-th metabolite is a substrate of the $j$-th reaction and $r_{j,m} = 0$ otherwise. Using the matrix $\mathbf{R}(\mathbf{r})$, the matrix $\mathbf{A}(\mathbf{r}) \in \mathbb{R}^{(J+L)\times(M+K)}$ is defined

as

$$\mathbf{A}(\mathbf{r}) \coloneqq \left(\begin{array}{c|c} \mathbf{R}(\mathbf{r}) & -\mathbf{C} \\ \hline -\mathbf{D}^{\mathrm{T}} & \mathbf{0}_{L\times K} \end{array}\right),$$

where $\mathbf{0}_{L\times K} \in \mathbb{R}^{L\times K}$ is a zero matrix, whose elements are all zero, $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_K) \in \mathbb{R}^{J\times K}$, $\mathbf{c}_k$ is the $k$-th basis of $\ker \boldsymbol{\nu}$, which indicates the right null space of $\boldsymbol{\nu}$, $K$ is the number of the bases of $\ker \boldsymbol{\nu}$, $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_L) \in \mathbb{R}^{M\times L}$, $\mathbf{d}_l$ is the $l$-th basis of $\ker \boldsymbol{\nu}^{\mathrm{T}}$, which indicates the right null space of $\boldsymbol{\nu}$, and $L$ is the number of the bases of $\ker \boldsymbol{\nu}^{\mathrm{T}}$. Note that $\mathbf{A}(\mathbf{r})$ is in general a square matrix, *i.e.*, $J + L = M + K$. A sensitivity matrix is defined as

$$\mathbf{S}(\mathbf{r}) \coloneqq -\mathbf{A}(\mathbf{r})^{-1}. \tag{1}$$

Here, the $(m, j)$ element of $\mathbf{S}(\mathbf{r}) \in \mathbb{R}^{(M+K)\times(J+L)}$ is proven to be equal to a constant multiple of a quantitative response of the $m$-th metabolite concentration/flux to a perturbation of the $j$-th reaction/conserved quantity [11]. Even though we calculate the sensitivity matrix, we cannot determine the quantitative response value. However, each element of the sensitivity matrix has the same sign (positive, negative, or zero) as the corresponding quantitative response value, enabling us to discuss qualitative responses. Although flux responses are represented as sets of reactions that are non-zero in $\mathbf{c}_k$, we can obtain a flux response corresponding to each reaction by calculating linear combinations. Let $\mathbf{T}(\mathbf{r}) \in \mathbb{R}^{J\times(J+L)}$ be a matrix indicating the responses of each reaction flux, which can be written as

$$\mathbf{T}(\mathbf{r}) \coloneqq \mathbf{C}\mathbf{S}(\mathbf{r})_{M+1:M+K,1:J+L},$$

where $\mathbf{S}(\mathbf{r})_{M+1:M+K,1:J+L} \in \mathbb{R}^{K\times(J+L)}$ is a block matrix of $\mathbf{S}(\mathbf{r})$, which is extracted from the $(M+1)$-th to $(M + K)$-th rows.

### 2.1.2 Qualitative response prediction

SSA can predict qualitative responses (positive, negative, zero, or indefinite) to perturbations using only structural information, which is the metabolic network and the constraints on the $\mathbf{r}$ values. Here, we describe this in a precise way.

Let $\mathbf{Q}(\mathbf{r}) \in \{-1, 0, 1\}^{(M+J)\times(J+L)}$ be the qualitative response matrix, which is given by

$$\mathbf{Q}(\mathbf{r}) \coloneqq \mathrm{sign}\left(\left(\begin{array}{c} \mathbf{S}(\mathbf{r})_{1:M,1:J+L} \\ \hline \mathbf{T}(\mathbf{r}) \end{array}\right)\right),$$

where sign$(\cdot)$ is the element-wise sign function, which returns a matrix whose element equals 1, 0, and -1 if the sign of the corresponding element of the given matrix is positive, zero, and negative, respectively. The responses of the $m$-th metabolite concentration and the $j$-th reaction flux to perturbations correspond to the $m$-th and $(M + j)$-th rows of $\mathbf{Q}(\mathbf{r})$, respectively. We refer to the $m$-th row and the $j$-th column of $\mathbf{Q}(\mathbf{r})$ as the $m$-th "observation target" and the $j$-th "perturbation target", respectively. In addition, we call the perturbation experiment/prediction/response for the $m$-th observation and $j$-th perturbation targets the $(m, j)$ experiment/prediction/response.

Since SSA is concerned with general results of qualitative responses, the elements of $\mathbf{Q}(\mathbf{r})$ are examined across all possible values of $\mathbf{r}$. It is important to note that there are typically constraints on the $\mathbf{r}$ values, such as $r_{j,m} > 0$, and $\mathbf{Q}(\mathbf{r})$ is evaluated within these constraints. Let $q_{m,j}(\mathbf{r})$ denote the $(m, j)$ element of $\mathbf{Q}(\mathbf{r})$ $(m = 1, \ldots, (M + J),\ j = 1, \ldots, (J + L))$. The qualitative response for each $q_{m,j}(\mathbf{r})$ can be classified into one of four categories; i) $q_{m,j}(\mathbf{r})$ is zero for any $\mathbf{r}$. This case is a consequence of structural properties, as explained by a theorem known as the law of localization and buffering structures [35, 11]. ii) $q_{m,j}(\mathbf{r})$ is positive for any $\mathbf{r}$. iii) $q_{m,j}(\mathbf{r})$ is negative for any $\mathbf{r}$. iv) $q_{m,j}(\mathbf{r})$ varies depending on the quantitative values of $\mathbf{r}$, making the sign of the response indefinite. For example, including a term $r_1 - r_2$ in the symbolic expression corresponding to $q_{m,j}(\mathbf{r})$ makes $q_{m,j}(\mathbf{r})$ positive if $r_1 > r_2$ and negative if $r_1 < r_2$, where $r_i$ is the $i$-th element of $\mathbf{r}$. These four categories are referred to as zero, positive, negative, and indefinite, respectively.

There are several methods for evaluating $q_{m,j}(\mathbf{r})$ for all possible $\mathbf{r}$. One approach is to perform symbolic calculations, regarding $\mathbf{r}$ as symbolic variables [10]. Although this method is rigorous, it is only practical for relatively small metabolic networks due to its computational complexity. An alternative, more computationally tractable method is to draw a sufficient number of $\mathbf{r}$ samples from an arbitrary probabilistic distribution and numerically evaluating $\mathbf{Q}(\mathbf{r})$ [36]. This latter approach has inspired us to introduce stochastic variables into SSA in this study.

# 3 Methods

## 3.1 Our model

In SSA, the predictive response is obtained by considering all cases of $\mathbf{r}$. $\mathbf{r}$ depends on $\{k_j\}_j$ and $\{x_m\}_m$, which, in turn, depend on environmental conditions, such as aerobic/anaerobic states, pH levels, and nutrient availability, as well as individual differences among microbes. The SSA qualitative response prediction described in the previous section thus focuses on the general consequences of responses drawn from structural information alone. We propose a Bayesian statistical model based on SSA, named BayesianSSA, to integrate environmental information extracted from perturbation data into SSA predictions. We consider a probability of each $\mathbf{r}$ value, regarding $\mathbf{r}$ as a stochastic variable. The BayesianSSA posterior probability of $\mathbf{r}$ reflects perturbation data and thus extracts environmental information. Predicting responses on the basis of the posterior distribution of $\mathbf{r}$ and positivity confidence values enables us to integrate the environmental information into SSA predictions. A positivity confidence value, which we proposed in this study, indicates the probability that the predictive response is positive. The positivity confidence values enable us to interpret indefinite predictions stochastically. We will describe the details in the "Positivity confidence value" section.

### 3.1.1 Prior distribution and likelihood

In this section, we describe the likelihood function and the prior distribution of BayesianSSA. Suppose that we have a perturbation dataset $\mathbf{y}$, which shows the signs of experimentally observed responses. We define the perturbation record, an element of $\mathbf{y}$, as follows:

$$y_i \coloneqq \begin{cases} 1 & \text{if increase} \\ -1 & \text{if decrease} \end{cases}, \tag{2}$$

where $y_i$ is the $i$-th perturbation record, obtained from the $(m_i, j_i)$ experiment, and "increase" and "decrease" indicate the cases where the observation target increases and decreases when the perturbation target is perturbed, respectively (see the "Qualitative response prediction" section for the definition of observation and perturbation targets). We consider the probability that $y_i \neq q_{m_i,j_i}(\mathbf{r})$ because experimental errors may occur even if $q_{m,j}(\mathbf{r})$ makes the correct prediction, assuming that the

likelihood function is the following:

$$p(y_i|\mathbf{r},\boldsymbol{\rho}) = \begin{cases} \rho_{m_i,j_i} & \text{if } y_i = q_{m_i,j_i}(\mathbf{r}) \\ 1-\rho_{m_i,j_i} & \text{if } y_i \neq q_{m_i,j_i}(\mathbf{r}) \end{cases}, \tag{3}$$

where $\rho_{m,j} \in (0,1)$ is a parameter indicating the reliability of the $(m,j)$ experiment, and $\boldsymbol{\rho}$ is a matrix whose $(m,j)$ element is $\rho_{m,j}$. The likelihood function means that the probability of $y_i$ is $\rho_{m_i,j_i}$ if $\mathbf{r}$ can accurately predict $y_i$, and is $1-\rho_{m_i,j_i}$ otherwise. In other words, BayesianSSA assumes that the result of the $i$-th experiment can stochastically vary in accordance with the probability of $\rho_{m_i,j_i}$. $\rho_{m,j}$ is different for each $(m,j)$, and each $(m,j)$ experiment is assumed to have different reliability in BayesianSSA. This assumption is reasonable because the distribution of measured values is supposed to be different for each $(m,j)$ experiment.

We consider the prior distribution of $\rho_{m,j}$ as

$$p(\rho_{m,j}) = \mathcal{B}\left(\rho_{m,j}|a,b\right),$$

where $\mathcal{B}\left(\cdot|a,b\right)$ denotes the beta distribution probability density function with parameters $a \in \mathbb{R}_{>0}$ and $b \in \mathbb{R}_{>0}$. The prior distribution of $\mathbf{r}$ should be chosen in accordance with the constraint on $\mathbf{r}$. A typical constraint on $r_{j,m}$ is $r_{j,m} > 0$ with the $m$-th metabolite being the substrate of the $j$-th reaction, and we consider only such type of constraints in this study. We used a weighted empirical distribution with samples drawn from a log-normal distribution as the prior distribution. Specifically, the prior distribution of $\mathbf{r}$ is the following:

$$\begin{aligned} p(\mathbf{r}=\mathbf{r}^{(v)}) &= \mathcal{WE}\left(\mathbf{r}=\mathbf{r}^{(v)}\middle|\mathbf{w}\right) \\ &= w_v, \end{aligned} \tag{4}$$

where $\mathcal{WE}\left(\mathbf{r}=\mathbf{r}^{(v)}\middle|\mathbf{w}\right)$ denotes the weighted empirical distribution probability mass function with a stochastic variable $\mathbf{r}$, a weight parameter $\mathbf{w} = (w_1,\ldots,w_V)^{\mathrm{T}}$, and a parameter sample set $\left\{\mathbf{r}^{(v)}\right\}_{v=1}^{V}$, $w_v > 0$ and $\sum_{v=1}^{V} w_v = 1$, and $V$ is the size of the sample set. Here, we generated the parameter sample set $\left\{\mathbf{r}^{(v)}\right\}_{v=1}^{V}$ as follows:

$$\mathbf{r}^{(v)} \sim \mathcal{LN}\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right),$$

where $\mathcal{LN}\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$ denotes the log-normal distribution with parameters $\boldsymbol{\mu} \in \mathbb{R}^P$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{P\times P}$ ($\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ correspond to the mean and covariance matrix parameter of the normal distribution, respectively), and $a \sim \mathcal{D}$ indicates that a stochastic variable $a$ is drawn from a distribution $\mathcal{D}$.

The purpose of BayesianSSA is to obtain a better distribution $p(\mathbf{r})$ for calculating positivity confidence values (Eq. (7)). Therefore, we need the marginal posterior distribution $p(\mathbf{r}|\mathbf{y})$. The marginal likelihood is obtained as

$$\begin{aligned} p(\mathbf{y}|\mathbf{r}=\mathbf{r}^{(v)}) &= \int p(\mathbf{y}|\mathbf{r}=\mathbf{r}^{(v)},\boldsymbol{\rho})p(\boldsymbol{\rho})d\boldsymbol{\rho} \\ &= \prod_{m=1}^{M+J}\prod_{j=1}^{J+L} \frac{\mathrm{Beta}\left(\hat{a}_{m,j,v},\hat{b}_{m,j,v}\right)}{\mathrm{Beta}(a,b)}, \end{aligned}$$

where $\mathrm{Beta}(\cdot,\cdot)$ is the beta function, $\hat{a}_{m,j,v} = t_{m,j,v}+a$, $\hat{b}_{m,j,v} = f_{m,j,v}+b$, and $t_{m,j,v}$ and $f_{m,j,v}$ are the number of true and false $(m,j)$ predictions based on $\mathbf{r}^{(v)}$ in $\mathbf{y}$, respectively. Here, $t_{m,j,v} = f_{m,j,v} = 0$ for a $(m,j)$ experiment that has not been conducted.

If a continuous prior and posterior distribution is desired, one can use them and obtain samples from the posterior distribution using the Markov chain Monte Carlo (MCMC) method. Since we discretized the log-normal distribution (Eq. (4)), we can calculate the posterior distribution without approximation using MCMC (details are described in the "Calculating posterior distribution" section).

### 3.1.2 Calculating posterior distribution

The marginalized posterior distribution $p(\mathbf{r}|\mathbf{y})$ in BayesianSSA can be calculated by normalizing the following formula:

$$\begin{aligned} p(\mathbf{r} = \mathbf{r}^{(v)}|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{r} = \mathbf{r}^{(v)})p(\mathbf{r} = \mathbf{r}^{(v)}) \\ &= \left( \prod_{m=1}^{M+J} \prod_{j=1}^{J+L} \frac{\mathrm{Beta}\left(\hat{a}_{m,j,v}, \hat{b}_{m,j,v}\right)}{\mathrm{Beta}(a, b)} \right) w_v, \end{aligned}$$

We can omit calculating the constant of this formula, and we derive that

$$p(\mathbf{r} = \mathbf{r}^{(v)}|\mathbf{y}) = \frac{g(\mathbf{y}, \mathbf{r}^{(v)})}{\sum_{v'=1}^{V} g(\mathbf{y}, \mathbf{r}^{(v')})}, \tag{5}$$

where

$$\begin{aligned} g(\mathbf{y}, \mathbf{r}^{(v)}) &\coloneqq w_v \prod_{(m,j)\in\Lambda} \mathrm{Beta}\left(\hat{a}_{m,j,v}, \hat{b}_{m,j,v}\right), \\ \Lambda &\coloneqq \{(m, j) \mid t_{m,j,v} + f_{m,j,v} \neq 0\}. \end{aligned}$$

We implemented these calculations using the log-sum-exp trick.

### 3.1.3 Calculating predictive distribution

To evaluate predictive performance, we calculate the predictive probability $p(\mathbf{y}^{\mathrm{new}}|\mathbf{y})$, where $\mathbf{y}^{\mathrm{new}}$ is a new sample that is not used in BayesianSSA fitting. Using the posterior distribution $p(\mathbf{r}, \boldsymbol{\rho}|\mathbf{y})$ (Supplementary Section S1), the predictive probability can be obtained as

$$\begin{aligned} p(\mathbf{y}^{\mathrm{new}}|\mathbf{y}) &= \sum_{v=1}^{V} \int p(\mathbf{y}^{\mathrm{new}}|\mathbf{r} = \mathbf{r}^{(v)}, \boldsymbol{\rho}) \\ &\qquad p(\mathbf{r} = \mathbf{r}^{(v)}, \boldsymbol{\rho}|\mathbf{y}) d\boldsymbol{\rho} \\ &= \frac{\sum_{v=1}^{V} w_v \prod_{(m,j)\in\Lambda^{\mathrm{new}}} \mathrm{Beta}\left(\hat{a}^{\mathrm{new}}_{m,j,v}, \hat{b}^{\mathrm{new}}_{m,j,v}\right)}{\sum_{v=1}^{V} w_v \prod_{(m,j)\in\Lambda^{\mathrm{new}}} \mathrm{Beta}\left(\hat{a}_{m,j,v}, \hat{b}_{m,j,v}\right)} \end{aligned} \tag{6}$$

where

$$\begin{aligned} \hat{a}^{\mathrm{new}}_{m,j,v} &\coloneqq \hat{a}_{m,j,v} + t^{\mathrm{new}}_{m,j,v}, \\ \hat{b}^{\mathrm{new}}_{m,j,v} &\coloneqq \hat{b}_{m,j,v} + f^{\mathrm{new}}_{m,j,v}, \\ \Lambda^{\mathrm{new}} &\coloneqq \Lambda \cup \left\{(m, j) \mid t^{\mathrm{new}}_{m,j,v} + f^{\mathrm{new}}_{m,j,v} \neq 0\right\}, \end{aligned}$$

and $t^{\mathrm{new}}_{m,j,v}$ and $f^{\mathrm{new}}_{m,j,v}$ are the number of true and false $(m, j)$ predictions based on $\mathbf{r}^{(v)}$ in $\mathbf{y}^{\mathrm{new}}$, respectively. The detailed derivation is described in Supplementary Section S2.

### 3.1.4 Bayesian updating

Bayesian updating, a procedure where the posterior distribution is used as a prior distribution for the next estimation, can be easily applied to BayesianSSA. In BayesianSSA, updating $\hat{a}_{m,j,v}$ and $\hat{b}_{m,j,v}$ to $\hat{a}^{\mathrm{new}}_{m,j,v}$ and $\hat{b}^{\mathrm{new}}_{m,j,v}$ is equivalent to Bayesian updating (Supplementary Section S3). This update is also derived from the fact that the final updated posterior distribution in Bayesian updating does not depend on the order of the given perturbation data due to Bayes' theorem [37].

#### 3.1.5 Positivity confidence value

The introduction of stochastic variables $\mathbf{r}$ enables interpreting indefinite predictions in SSA. We define positivity confidence values as the probabilities that the responses are positive for each indefinite prediction. The positivity confidence value of the $(m, j)$ prediction for a distribution $p^\star(\mathbf{r})$ is written as

$$c_p(m, j) \coloneqq \mathbb{E}_{p^\star(\mathbf{r})}[\mathbb{I}_{q_{m,j}(\mathbf{r})=1}], \tag{7}$$

where $\mathbb{E}_{p^\star(\mathbf{r})}[\cdot]$ denotes the expectation with respect to $p^\star(\mathbf{r})$, and $\mathbb{I}_{a=b}$ is an indicator that is equal to 1 if $a = b$ and 0 otherwise. A higher positivity confidence value indicates that the qualitative response is more likely to be positive, and the probability of being negative is higher than that of being positive when the positivity confidence value is below 0.5. Note that we use the predictive distributions derived in the "Calculating predictive distribution" section rather than positivity confidence values to evaluate predictive performance on perturbation datasets. This is because positivity confidence values do not consider experimental error.

### 3.2 Used data

#### 3.2.1 Metabolic network information

We utilized the central metabolic pathway of *E. coli* MG1655 used in a previous study [38]. This metabolic network was originally from the EcoCyc database [39] and modified by Trinh *et al.* [40] and Toya *et al.* [38]. We preprocessed this dataset as follows:

1. Remove the biomass objective function.
2. Remove metabolites that have no reactions that produce or use them.
3. Remove reactions that no longer have substrates or products due to the previous processes.
4. Integrate cytoplasm and extracellular metabolites.

The second step is necessary to apply SSA to the network because it is typically impossible to calculate the inverse of the matrix $\mathbf{A}(\mathbf{r})$ (Eq. (1)) when metabolites not involved in the flow are included in the network. Here, metabolites in flow refer to metabolites that serve as both inputs and outputs of reactions included in the network. The fourth step aims to reduce computation time, and this procedure does not alter the results from SSA. After these preprocessing steps, we converted the resulting network into the stoichiometric matrix $\boldsymbol{\nu}$. Constraints on $\mathbf{R}(\mathbf{r})$ were determined solely by the metabolic network, where we set $r_{j,m} > 0$ if the $j$-th metabolite is a substrate of the $m$-th reaction and $r_{j,m} = 0$ otherwise. The resulting metabolite and reaction lists are shown in Supplementary Table S2 and Supplementary Table S3. We use the abbreviations in Supplementary Table S2 and Supplementary Table S3 in the following.

#### 3.2.2 Synthetic data generation

We generated synthetic data to compare BayesianSSA with random and base methods. The synthetic perturbation dataset is generated in accordance with a Bernoulli distribution with parameters generated in accordance with beta distributions. Then, we replaced all 0 with $-1$ to match the support of the Bernoulli distribution and the range of the perturbation record $y_i$ (Eq. (2)). We used GND, PTS, and PPC as the perturbation targets and succinate export (SUCCt) as the observation target. We used a beta distribution with an expectation is close to 0 for GND and PTS and a beta distribution with an expectation is close to 1 for PPC. These distribution settings were in accordance with previous reports [41, 42, 43, 44]. We used $(a, b) = (0.1, 0.3)$ and $(a, b) = (0.3, 0.1)$ as parameters of beta distributions with expectations are close to 0 and 1, respectively. Note that these parameters are used

only for the data generation and different from the parameters of the prior distributions, which are weakly informative. We used 30 as the number of records for each perturbation target. The obtained synthetic dataset is shown in Table 1.

Table 1: Obtained synthetic perturbation dataset

| Observation target | Perturbation target | Number of $i$ where $y_i = 1$ | Number of $i$ where $y_i = -1$ |
|---|---|---|---|
| SUCCt | GND | 0 | 30 |
| SUCCt | PPC | 25 | 5 |
| SUCCt | PTS | 0 | 30 |

### 3.2.3 Wet lab experiments

We conducted perturbation experiments for the reactions CS, FBP, ICL, LDH, ME1, ME2, PCK, PPC, PPS, and PTA. The genes corresponding to the reactions, listed in Table 2, were introduced into the plasmid vector pLEAD5 (NIPPON GENE CO., LTD.). We amplified the gene sequences using PrimeSTAR GXL DNA Polymerase (Takara Bio Inc.) with *E. coli* DH5$\alpha$ (NIPPON GENE CO., LTD.) as a template sequence. The primer sequences were derived from NCBI Genes [45]. The nucleotide sequences of DNAs were analyzed using the BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) with SeqStudio™ Genetic Analyzer (Applied Biosystems). The nucleotide sequence data were processed using GENETYX-Mac NETWORK software, version 15 (GENETYX CORPORATION). We introduced the constructed plasmid vectors into *E. coli* JM109 (NIPPON GENE CO., LTD.). The modified strains were aerobically cultured and anaerobically fermented at 37°C in M9 minimal medium (Thermo Fisher Scientific Inc.). To measure succinate concentrations, we used EnzyChrom Succinate Assay Kit (#ESNT-100, BioAssay Systems) and the absorbance meter SH-8000(CORONA ELECTRIC Co.,Ltd.) at 570 nm. The resulting absorbance, which is a constant multiple of the succinate concentrations, was normalized by the optical density of the bacterial liquid. We calculated the differences between the resulting values and the absorbances of controls, which are of a strain with only pLEAD5, and the signs of the values were used as the real perturbation data. We constructed and measured three replicates for each perturbed reaction. The obtained dataset is shown in Table 2. To validate the overexpression, we performed real-time PCR using QuantStudio5 Real-time PCR system (Thermo Fisher SCIENTIFIC) and confirmed the genes were successfully overexpressed (Supplementary Table S4).

Table 2: Obtained real perturbation dataset

| Observation target | Perturbation target | Gene of perturbation target | Number of $i$ where $y_i = 1$ | Number of $i$ where $y_i = -1$ |
|---|---|---|---|---|
| SUCCt | CS | gltA | 2 | 1 |
| SUCCt | FBP | fbp | 0 | 3 |
| SUCCt | ICL | aceA | 3 | 0 |
| SUCCt | LDH | ldhA | 0 | 3 |
| SUCCt | ME1 | maeA | 0 | 3 |
| SUCCt | ME2 | maeB | 0 | 3 |
| SUCCt | PCK | pckA | 3 | 0 |
| SUCCt | PPC | ppc | 0 | 3 |
| SUCCt | PPS | pps | 1 | 2 |
| SUCCt | PTA | pta | 1 | 2 |

## 3.3 Evaluation

### 3.3.1 Naive Bayes model

To evaluate the effectiveness of incorporating SSA, we constructed a naive Bayes model as follows:

$$p(y_i|\boldsymbol{\eta}) = \begin{cases} \eta_{m_i,j_i} & \text{if } y_i = 1 \\ 1 - \eta_{m_i,j_i} & \text{if } y_i \neq 1 \end{cases},$$

where $\eta_{m,j} \in (0,1)$ is a parameter indicating the probability that $y_i = 1$ where the $i$-th experiment is of $(m,j)$, and $\boldsymbol{\eta}$ is a matrix whose $(m,j)$ element is $\eta_{m,j}$. The predictive distribution of one new perturbation record $y^{\text{new}}$ is as follows:

$$p(y^{\text{new}}|\mathbf{y}) = \begin{cases} \frac{n^{(\text{p})}_{m,j}+1}{n^{(\text{p})}_{m,j}+n^{(\text{n})}_{m,j}+2} & \text{if } y^{\text{new}} = 1 \\ \frac{n^{(\text{n})}_{m,j}+1}{n^{(\text{p})}_{m,j}+n^{(\text{n})}_{m,j}+2} & \text{if } y^{\text{new}} \neq 1 \end{cases}, \tag{8}$$

where $n^{(\text{p})}_{m,j}$ and $n^{(\text{n})}_{m,j}$ are the numbers of the $(m,j)$ experiments in $\mathbf{y}$ where $y_i = 1$ and $y_i = -1$, respectively. We used $\mathcal{B}(\eta_{m,j}|1,1)$ as the prior distribution.

### 3.3.2 Cross-entropy loss

To evaluate the performance of BayesianSSA, we used the cross-entropy loss as follows:

$$\text{CE}(N) = -\sum_{i=1}^{N} \log p_i(y^{\text{new}} = y_i),$$

where $N$ is the number of trials, which means how many times data are added, and $p_i(\cdot)$ is the $i$-th predictive distribution of BayesianSSA or the naive Bayes model. The $i$-th predictive distribution is calculated using $\{y_{i'}\}_{i'=1}^{i-1}$. We used $p_t(y^{\text{new}} = 1) = 0.5$ and $p_t(y^{\text{new}} = y_i) = p_1(y^{\text{new}} = y_i)$ as the "random method" and "base method" to calculate the cross-entropy loss, respectively, to be compared with BayesianSSA. Here, $p_1(y^{\text{new}} = y_i)$ indicates the initial distribution of BayesianSSA.

# 4 Results

$m_{\text{name}}$ denotes the index of the observation target in the following. In other words, the $m_{\text{name}}$-th observation target is a metabolite or a reaction whose name is "name." Similarly, $j_{\text{name}}$ denotes the index of the perturbation target "name" in the following. Unless otherwise stated, we used $V = 10000$, $w_v = \frac{1}{V}$, $\boldsymbol{\mu} = \mathbf{0}_P$, $\boldsymbol{\Sigma} = \mathbf{I}_{P\times P}$, and $(a,b) = (3,1)$ as the hyper-parameters of BayesianSSA where $\mathbf{0}_P$ is the $P$-dimensional zero vector, and $\mathbf{I}_{P\times P}$ is the $P \times P$ identity matrix.

## 4.1 Performance evaluation on synthetic dataset

To compare predictive performance between BayesianSSA and the base method under an ideal situation that the experimental error is low, we examined the cross-entropy loss $\text{CE}(N)$ (*cf.* the "Cross-entropy loss" section) on the synthetic dataset shown in Table 1. Figure 1 shows the cross-entropy loss trajectory for each method. While cross-entropy loss values at the last trial of BayesianSSA with $(a,b) = (9,1)$, $(a,b) = (6,2)$, $(a,b) = (3,1)$, and $(a,b) = (2,1)$ are 19.6, 23.3, 21.4, and 22.3, respectively, those of the random and base methods are 62.4 and 66.6, respectively. BayesianSSA outperformed the random and base method from the perspective of prediction accuracy under the ideal situation.

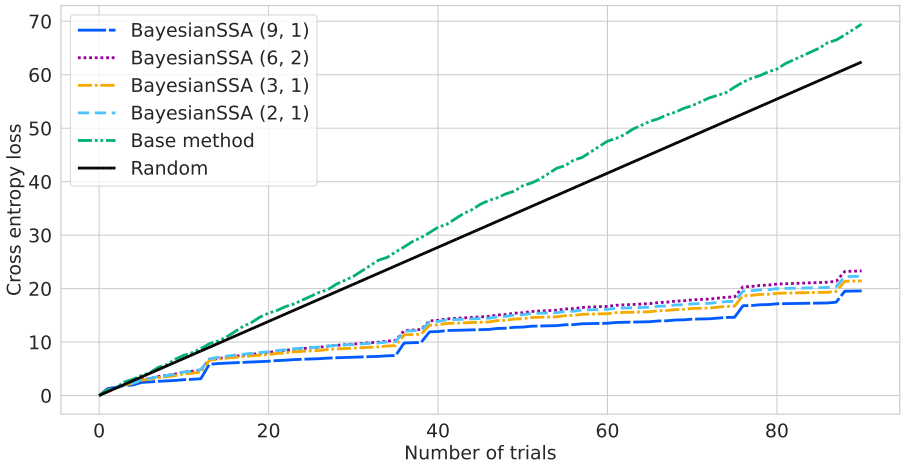


Figure 1: Cross-entropy loss trajectory for each method on the synthetic dataset. The $x$- and $y$-axes indicate the number of trials and cross-entropy loss, respectively.

## 4.2 Positivity confidence value transition on pseudo data

To examine the transition of positivity confidence values, we applied BayesianSSA to a pseudo perturbation dataset (shown in Table 3) based on the previous studies [41, 42, 43]. Figure 2 shows

Table 3: Pseudo perturbation dataset

| Observation target | Perturbation target | $y_i$ | The number of records |
|---|---|---|---|
| SUCCt | GND | $-1$ | 10 |
| SUCCt | PTS | $-1$ | 10 |

the transition of positivity confidence values for increasing SUCCt, *i.e.*, $c_p(m_{\mathrm{SUCCt}}, \cdot)$, when fitting BayesianSSA to the pseudo perturbation dataset. We found that the positivity confidence values were updated for multiple reactions rather than one reaction. The first update (from Figures 2(a) to 2(b)) shows the EDA (6PG $\rightarrow$ G3P + PYR) positivity confidence value for the production of succinate becomes higher. The second update (from Figure 2(b) to Figure 2(c)) shows the reactions included by the flux from FBP to PEP become lower. One of the reaction candidates to increase the SUCCt flux was PPC, whose reaction formula was PEP + CO2 $\rightarrow$ OAA. While the initial PPC positivity confidence value was 0.714 (Supplementary Table S6), the updated PPC positivity confidence value was 0.869 (Supplementary Table S7). A previous report showed the succinate production of *E. coli* increases when PPC was up-regulated [44], and the PPC positivity confidence value estimated by BayesianSSA is consistent with the report. Similarly, those of GND and PTS were updated from 0.530 and 0.779 to $7.84 \times 10^{-3}$ and $3.23 \times 10^{-2}$, respectively. Although these results are also consistent with the reports [41, 42, 43], they are naive because the information of these reports was used directly to fit BayesianSSA. The initial positivity confidence values, which were calculated on the basis of the prior distribution, and positivity confidence values updated with the pseudo perturbation dataset are shown in Supplementary Table S6 and S7, respectively.

## 4.3 Performance evaluation on real data

To compare the performances between BayesianSSA and the naive Bayes model, which uses only data without SSA, we examined the cross-entropy loss CE($N$) (*cf.* the "Cross-entropy loss" section). Figure 3 shows the cross-entropy loss trajectory for each method. The perturbation records were

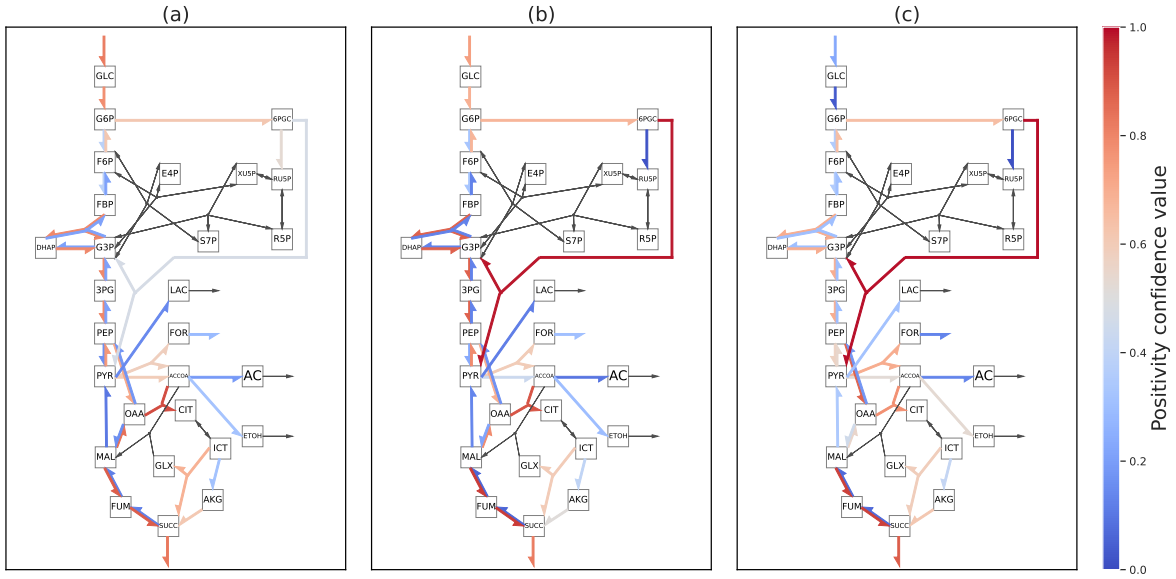


Figure 2: Transition of positivity confidence value for increasing the succinate export flux $c_p(m_{\mathrm{SUCCt}}, \cdot)$. Each square indicates a metabolite in the network. Each arrow indicates a reaction, and its color shows the positivity confidence value (red and blue) or zero response (grey). The reactions corresponding to the edges in this figure are shown in Supplementary Table S5. **(a)** Values given by the initial model. **(b)** Values given by the model updated by 10 perturbation records that $y_i = -1$ where $m_i = m_{\mathrm{SUCCt}}$ and $j_i = j_{\mathrm{GND}}$. **(c)** Values given by the model updated by 10 perturbation records that $y_i = -1$ where $m_i = m_{\mathrm{SUCCt}}$ and $j_i = j_{\mathrm{PTS}}$ in addition to the perturbation records of **(b)**.

randomly shuffled and used in each trial. The loss of BayesianSSA is comparable to that of the base method in the early trials but smaller in the late trials. While the cross-entropy loss values at the last trial of BayesianSSA with $(a, b) = (9, 1)$, $(a, b) = (6, 2)$, $(a, b) = (3, 1)$, and $(a, b) = (2, 1)$ are 14.8, 15.7, 15.4, and 16.1, respectively, those of the random method, the base method, and the naive Bayes model are 20.8, 19.0, and 17.2, respectively. BayesianSSA outperformed the random and base methods and the naive Bayes model from the perspective of prediction accuracy. While BayesianSSA made a prediction on the basis of the perturbation dataset and SSA, the prediction of the base method is only based on SSA. This result suggests that BayesianSSA can integrate environmental information of the real dataset into SSA predictions. Similarly, the difference between BayesianSSA and the naive Bayes is whether or not SSA is incorporated. This result also indicates the practicability of BayesianSSA and that incorporating SSA into statistical models improves predictive performance.

The main difference between BayesianSSA and the naive Bayes model is in the predictions for out-of-sample perturbations, which are of new $(m, j)$ experiments. To calculate the predictive distribution of a $(m, j)$ experiment, the naive Bayes model uses only the results of $(m, j)$ experiments (Eq. (8)) while BayesianSSA uses the results of all the $(m, j) \in \Lambda$ experiments (Eq. (6)). That is, BayesianSSA can leverage data to predict the responses to out-of-sample perturbations through the $\mathbf{r}$ posterior distribution. To validate the predictive performance for out-of-sample perturbations, we examined the predictive probabilities by splitting the real dataset. For example, the predictive probabilities of the three replicates for the reaction CS were calculated by fitting BayesianSSA to the real dataset except for CS. Figure 4 shows the distribution of predictive probabilities calculated by BayesianSSA for out-of-sample perturbations. Here, each replicate was evaluated separately. The median of the distribution was 0.67, which is better than random (0.5), and this result indicates that fitting BayesianSSA contributes to the predictive performance for out-of-sample perturbations.

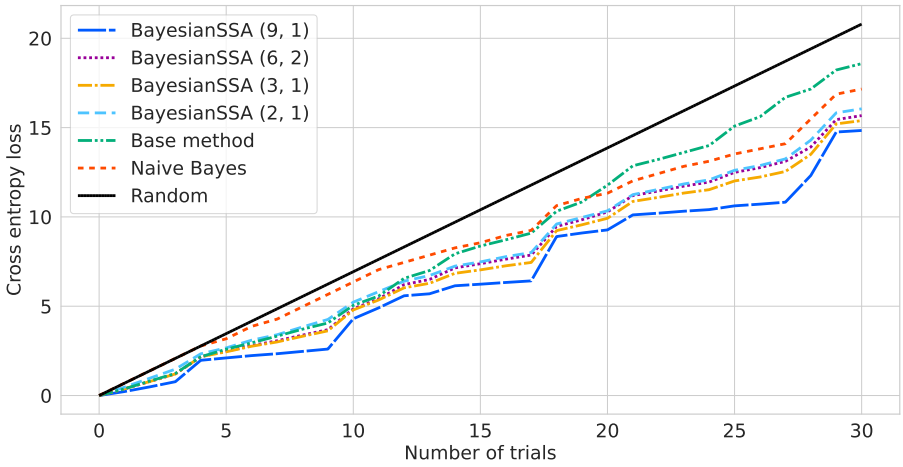


Figure 3: Cross-entropy loss trajectory for each method on the real dataset. The $x$- and $y$-axes indicate the number of trials and cross-entropy loss, respectively.

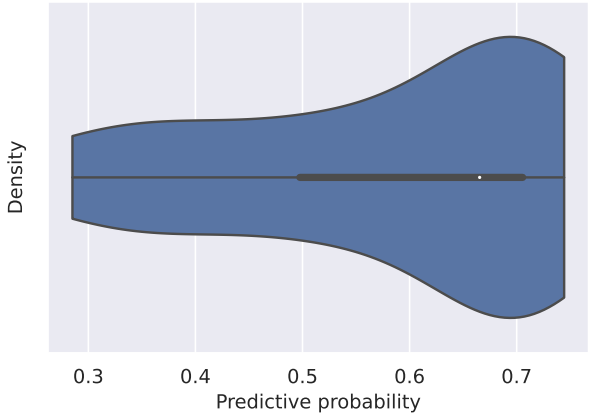


Figure 4: Distribution of predictive probabilities calculated by BayesianSSA for out-of-sample perturbations. Each value was calculated for the real dataset for new $(m, j)$ perturbation. The white dot indicates the median of the distribution.

## 4.4 Positivity confidence values on real data

To interpret the BayesianSSA estimation results, we examined the positivity confidence values after BayesianSSA was fitted to the real data whose observation target is the succinate export flux. Figure 5 shows the positivity confidence values in the metabolic network. All positivity confidence values updated by the real perturbation dataset are shown in Supplementary Table S8. We found high positivity confidence values of THD_r and NDH ($c_p(m_{\mathrm{SUCCt}}, j_{\mathrm{THD\_r}}) = 0.95$, $c_p(m_{\mathrm{SUCCt}}, j_{\mathrm{NDH}}) = 0.92$), which both use NADH. NDH produces Q8H2, which is required for FRD. In the tricarboxylic acid (TCA) cycle, ICL, MDH, FRD, FUM_r, and CS show high positivity confidence values ($> 0.85$). We also applied BayesianSSA with another prior distribution of $\mathbf{r}$, which is based on log-normal distributions with random parameters, to the real dataset (Supplementary Figure S1). All positivity confidence values using this prior distribution are shown in Supplementary Table S9. The reactions with $c_p(m, j) > 0.85$ are almost shared in the two BayesianSSA results. The only difference is the presence or absence of NDH (*cf.* Supplementary Table S8 and Supplementary Table S9). These results

suggest that BayesianSSA is robust to changes in the prior distributions of $\mathbf{r}$. We have also examined other observation targets besides succinate export flux and found that they can also be updated to high positivity confidence values (Supplementary Figure S2 and S3).

Positivity confidence values calculated by the BayesianSSA posterior distribution were consistent with previous reports. The positivity confidence values of CS and ICL, which are included in the glyoxylate pathway, were high. The glyoxylate pathway was reported as an essential pathway for succinate production [46, 47]. Similarly, the reductive pathway, which includes FUM_r and FRD, was reported as another essential pathway [46, 47]. Despite these consistencies, other previous reports for several reactions, such as PPC [44] and PTS [42, 43] are inconsistent with our results.

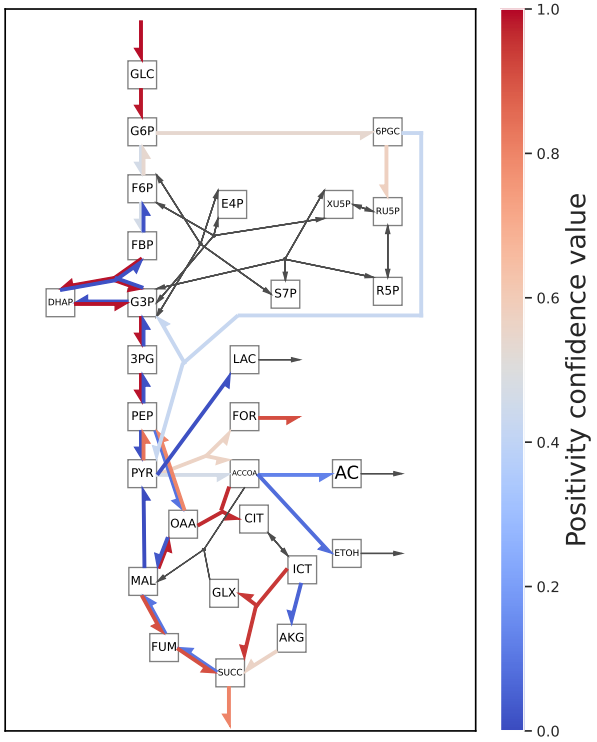


Figure 5: Positivity confidence values by BayesianSSA to increase the succinate export flux on metabolic networks. The positivity confidence values were calculated by BayesianSSA fitted to real data. Each square indicates a metabolite in the network. Each arrow indicates a reaction, and its color shows the positivity confidence value (red and blue) or zero response (grey). The reactions corresponding to the edges in this figure are shown in Supplementary Table S5.

## 4.5 Posterior distribution of r

To examine the differences between the prior and posterior distribution of $\mathbf{r}$, we compared $p(\mathbf{r}|\mathbf{y})$ with $p(\mathbf{r})$. Figure 6 shows $p(\mathbf{r})$ and $p(\mathbf{r}|\mathbf{y})$ on the first and second principal components of $\left\{\mathbf{r}^{(v)}\right\}_{v=1}^{V}$. We found that the posterior distribution had several peaks where the prior distribution only had one peak. This result indicates that the distribution of $\mathbf{r}$ is tailored to several cases of environments in which the used dataset was collected.

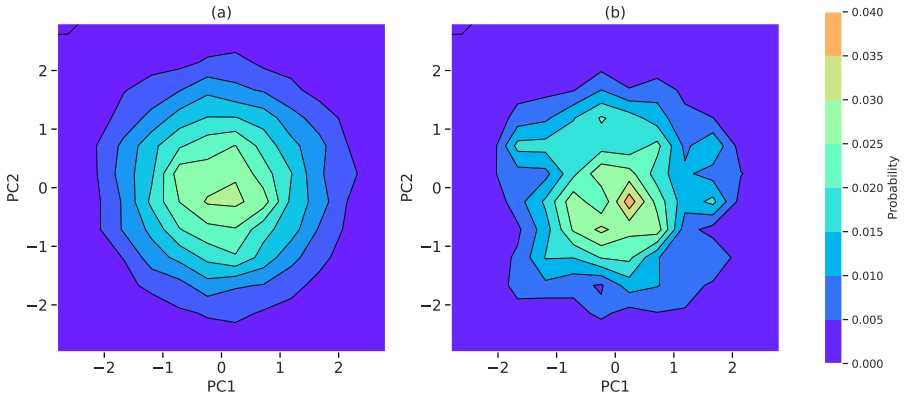


Figure 6: Contour plots of the prior **(a)** and posterior **(b)** distributions of $\log \mathbf{r}$. The $x$- and $y$-axes indicate the first and second principal components, respectively.

# 5 Discussion

We demonstrated the effectiveness of BayesianSSA on the basis of predictive performance using the real perturbation dataset where the observation target is the succinate export flux (SUCCt). BayesianSSA outperformed the base method (Figure 3), which is not fitted to the dataset, and this result showed BayesianSSA could integrate environmental information into SSA predictions. For validation of practicability, BayesianSSA also outperformed the random method and the naive Bayes model, and the performance of BayesianSSA for out-of-sample perturbations, which are of new $(m, j)$ experiments, was better than random. These results show that BayesianSSA can consider the relationships between different $(m, j)$ perturbations through $\mathbf{r}$, and that this consideration contributes to predictive performance.

We considered $\rho_{m,j}$ as a parameter and set the prior distribution in this study. As another option, $\rho_{m,j}$ can be given by an error rate of the measurement equipment used for the perturbation experiment. For example, consider a case where the error distribution of the used measurement equipment is a normal distribution with a mean parameter that equals the true value and a variance parameter $\sigma^2 = 1$ as an error distribution. If the experimental value obtained by the perturbation is 1, the probability that the true value is less than zero is approximately 16%. Therefore, setting $1 - \rho_{m,j} = 0.16$ can make BayesianSSA consider the error distribution of the measurement equipment. In this way, we can set a certain value of $\rho_{m,j}$. Note that we can easily calculate the posterior distribution of this model (Supplementary Section S4).

There are two directions for future work related to $\mathbf{r}$. First, the updated $p(\mathbf{r})$ may be used for response predictions in another metabolic network. When the reaction rate function $F_j$ and the probability distribution of $\mathbf{x}$ are equal between the two metabolic networks, $p(r_{j,m})$ can be used in another system that includes the $i$-th reaction and the $m$-th metabolite. Second, as previously discussed [10], we can consider allosteric regulation. Allosteric regulation is a type of regulation that increases/decreases reaction rates as a metabolite concentration increases [48, 49]. We can easily consider allosteric regulation by setting $r_{j,m} \neq 0$. Technically, $r_{j,m} < 0$ can be implemented by reversing the sign of $r_{j,m}$ after sampling $\mathbf{r}^{(v)}$. However, we need to know the $(j, m)$ pairs that have allosteric regulation in advance, and we omitted considering allosteric regulations in this study.

There is room for choice regarding the prior distribution of $\mathbf{r}$. First, continuous distributions can be adopted. We used the empirical distribution with samples from log-normal distributions as the $\mathbf{r}$ prior distribution for all experiments (*cf.* Eq. (4)). As long as the constraint on $\mathbf{r}$ is satisfied, other distributions can be adopted. However, the likelihood function changes discretely (*cf.* Eq. (3)), and the advantage of adopting a continuous distribution with employing MCMC methods may be limited. Second, ensemble approaches for several types of prior distributions may be effective. As shown in

Figure 5 and Supplementary Figure S1, the effect of the prior distribution is not negligible when dealing with a limited sample size. Using several types of prior distributions may contribute to making robust predictions.

Although we omitted the biomass production processes in this study, considering them can improve the representation of real biological activity. One commonly used approach in FBA involves optimizing the biomass objective function [18, 19]. However, since the biomass objective function depends on the specific strains and environmental conditions [50], it is difficult to use the biomass objective function when analyzing unfamiliar strains. Another difficulty is that the biomass objective function is a pseudo-reaction that contains multiple reactions and cannot be treated kinetically. Unlike FBA-based methods, which can consider biomass production processes as a single reaction, kinetics-based methods including BayesianSSA need to faithfully model the biomass production process. That is, it is necessary to define the biomass production rate equations as in a previous study [21].

Utilizing the positivity confidence value calculation (the "Positivity confidence value" section) and Bayesian updating (the "Bayesian updating" section) in BayesianSSA, we can construct an iterative design-build-test-learn (DBTL) cycle [51] on the basis of BayesianSSA for proposals of reactions to be perturbed. Specifically, the procedure is as follows:

1. Calculate positivity confidence values by Eq. (7).
2. Obtain a proposal of which perturbation and observation targets are validated in accordance with positivity confidence values.
3. Conduct perturbation experiments in accordance with the proposal.
4. Update the posterior distributions by Eq. (5).
5. Return to the first step.

One advantage of this scheme is the high efficiency because the experimental validation of proposals obtained by BayesianSSA is also a process collecting data for updating the BayesianSSA posterior distribution.

# 6 Conclusions

In this study, we proposed BayesianSSA, a Bayesian statistical model based on SSA. SSA was previously developed as a method to predict qualitative responses to enzyme perturbations on the basis of the structural information of the reaction network. However, the network structural information can sometimes be insufficient to predict qualitative responses unambiguously, which is a practical issue in bioproduction applications. To address this, BayesianSSA extracts environmental information from perturbation datasets collected in environments of interest and integrates it into SSA predictions. We applied BayesianSSA to synthetic and real datasets of the central metabolic pathway of *E. coli*. As a result, BayesianSSA outperformed the base method, which is the same as the BayesianSSA model but utilizes an initial prior distribution without incorporating perturbation datasets. This result shows that BayesianSSA can successfully integrate environmental information extracted from perturbation data into SSA predictions. In addition, the positivity confidence values estimated by BayesianSSA for increasing the succinate export flux were consistent with the known pathways reported to enhance the flux in previous studies. We believe that BayesianSSA will accelerate the chemical bioproduction process and contribute to advancements in the field.

# Declarations

**Ethics approval and consent to participate** Not applicable.

**Consent for publication** Not applicable.

**Availability of data and materials** Supplementary material is available from the journal website. The implementation of the algorithm is available on GitHub (`https://github.com/shion-hosoda-hitachi/BayesianSSA`).

**Competing interests** Not applicable.

**Funding** Not applicable.

**Authors' contributions** S.H., T.O., A.M., and M.S. conceptualized this study. S.H. devised the model, designed the algorithms, implemented the software, and performed all the computational experiments. H.I. performed all the wet lab experiments. T.M., H.I., and M.T. established the wet lab experimental procedures. S.H., M.S., T.O., and A.M. interpreted the computational results. S.H. and M.S. investigated previous researches for biological insights. S.H. wrote the draft. T.O., A.M., M.S., M.T., H.I., and T.M. revised the manuscript critically. All authors read and approved the final manuscript.

**Acknowledgements** We would like to thank Y. Mizunuma for her technical assistance with the wet lab experiments. We also thank Dr. K. Yokoyama for technical guidance. We appreciate the advice and technical support with the wet lab experiments from Dr. T. Takeya. We are grateful to Dr. A. Kandori, Dr. K. Watanabe, and Dr. S. Yabuuchi for their guidance throughout our project. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

# References

[1] Ólafur Ögmundarson, Sumesh Sukumara, Markus J. Herrgård, and Peter Fantke. Combining Environmental and Economic Performance for Bioprocess Optimization. *Trends in Biotechnology*, 38(11):1203–1214, November 2020.

[2] Apostolos Tsopanoglou and Ioscani Jiménez del Val. Moving towards an era of hybrid modelling: Advantages and challenges of coupling mechanistic and data-driven models for upstream pharmaceutical bioprocesses. *Current Opinion in Chemical Engineering*, 32:100691, June 2021.

[3] Anurag S. Rathore, Somesh Mishra, Saxena Nikita, and Priyanka Priyanka. Bioprocess Control: Current Progress and Future Perspectives. *Life*, 11(6):557, June 2021.

[4] Bunushree Behera, Yuwalee Unpaprom, Rameshprabu Ramaraj, Gaanty Pragas Maniam, Natanamurugaraj Govindan, and Balasubramanian Paramasivan. Integrated biomolecular and bioprocess engineering strategies for enhancing the lipid yield from microalgae. *Renewable and Sustainable Energy Reviews*, 148:111270, September 2021.

[5] Christopher A. Voigt. Synthetic biology 2020–2030: Six commercially-available products that are changing our world. *Nature Communications*, 11(1):6379, December 2020.

[6] Jay Keasling, Hector Garcia Martin, Taek Soon Lee, Aindrila Mukhopadhyay, Steven W. Singer, and Eric Sundstrom. Microbial production of advanced biofuels. *Nature Reviews Microbiology*, 19(11):701–715, November 2021.

[7] Xiaoyan Zhuang, Yonghui Zhang, An-Feng Xiao, Aihui Zhang, and Baishan Fang. Applications of Synthetic Biotechnology on Carbon Neutrality Research: A Review on Electrically Driven Microbial and Enzyme Engineering. *Frontiers in Bioengineering and Biotechnology*, 10, 2022.

[8] Lin Wang, Satyakam Dash, Chiam Yu Ng, and Costas D. Maranas. A review of computational tools for design and reconstruction of metabolic pathways. *Synthetic and Systems Biotechnology*, 2(4):243–252, December 2017.

[9] Anastasia Sveshnikova, Homa MohammadiPeyhani, and Vassily Hatzimanikatis. Computational tools and resources for designing new pathways to small molecules. *Current Opinion in Biotechnology*, 76:102722, August 2022.

[10] Atsushi Mochizuki and Bernold Fiedler. Sensitivity of chemical reaction networks: A structural approach. 1. Examples and the carbon metabolic network. *Journal of Theoretical Biology*, 367:189–202, February 2015.

[11] Takashi Okada and Atsushi Mochizuki. Sensitivity and network topology in chemical reaction systems. *Physical Review E*, 96(2):022322, August 2017.

[12] Amanda K. Fisher, Benjamin G. Freedman, David R. Bevan, and Ryan S. Senger. A review of metabolic and enzymatic engineering strategies for designing and optimizing performance of microbial cell factories. *Computational and Structural Biotechnology Journal*, 11(18):91–99, August 2014.

[13] Weihua Guo, Jiayuan Sheng, and Xueyang Feng. Mini-review: In vitro Metabolic Engineering for Biomanufacturing of High-value Products. *Computational and Structural Biotechnology Journal*, 15:161–167, January 2017.

[14] Priti Pharkya and Costas D. Maranas. An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metabolic Engineering*, 8(1):1–13, January 2006.

[15] Sridhar Ranganathan, Patrick F. Suthers, and Costas D. Maranas. OptForce: An Optimization Procedure for Identifying All Genetic Manipulations Leading to Targeted Overproductions. *PLOS Computational Biology*, 6(4):e1000744, April 2010.

[16] Michael J. McAnulty, Jiun Y. Yen, Benjamin G. Freedman, and Ryan S. Senger. Genome-scale modeling using flux ratio constraints to enable metabolic engineering of clostridial metabolism in silico. *BMC Systems Biology*, 6(1):42, May 2012.

[17] Shouyong Jiang, Irene Otero-Muras, Julio R. Banga, Yong Wang, Marcus Kaiser, and Natalio Krasnogor. OptDesign: Identifying Optimum Design Strategies in Strain Engineering for Biochemical Production. *ACS Synthetic Biology*, 11(4):1531–1541, April 2022.

[18] Jeffrey D. Orth, Ines Thiele, and Bernhard Ø. Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–248, March 2010.

[19] Adam M. Feist and Bernhard O. Palsson. The Biomass Objective Function. *Current opinion in microbiology*, 13(3):344–349, June 2010.

[20] Yoshihiro Usuda, Yosuke Nishio, Shintaro Iwatani, Stephen J. Van Dien, Akira Imaizumi, Kazutaka Shimbo, Naoko Kageyama, Daigo Iwahata, Hiroshi Miyano, and Kazuhiko Matsui. Dynamic modeling of Escherichia coli metabolic and regulatory systems for amino-acid production. *Journal of Biotechnology*, 147(1):17–30, May 2010.

[21] Hiroyuki Kurata and Yurie Sugimoto. Improved kinetic model of *Escherichia coli* central carbon metabolism in batch and continuous cultures. *Journal of Bioscience and Bioengineering*, 125(2):251–257, February 2018.

[22] Mohammadreza Yasemi and Mario Jolicoeur. Modelling Cell Metabolism: A Review on Constraint-Based Steady-State and Kinetic Approaches. *Processes*, 9(2):322, February 2021.

[23] Atsushi Mochizuki. A structural approach to understanding enzymatic regulation of chemical reaction networks. *The Biochemical Journal*, 479(11):1265–1283, June 2022.

[24] Jimena Di Maggio, Juan C. Diaz Ricci, and M. Soledad Diaz. Global Sensitivity Analysis in dynamic metabolic networks. In Jacek Jeżowski and Jan Thullie, editors, *Computer Aided Chemical Engineering*, volume 26 of *19 European Symposium on Computer Aided Process Engineering*, pages 1075–1080. Elsevier, January 2009.

[25] Henrik Kacser. The control of flux. In *Symp. Soc. Exp. Biol.*, volume 28, pages 65–101, 1973.

[26] Mary C. Wildermuth. Metabolic control analysis: Biological applications and insights. *Genome Biology*, 1(6):reviews1031.1, December 2000.

[27] Matthew L. Rizk and James C. Liao. Ensemble Modeling for Aromatic Production in Escherichia coli. *PLOS ONE*, 4(9):e6903, September 2009.

[28] Hiroki Nishiguchi, Natsuki Hiasa, Kiyoka Uebayashi, James Liao, Hiroshi Shimizu, and Fumio Matsuda. Transomics data-driven, ensemble kinetic modeling for system-level understanding and engineering of the cyanobacteria central metabolism. *Metabolic Engineering*, 52:273–283, March 2019.

[29] Jonathan Strutz, Jacob Martin, Jennifer Greene, Linda Broadbelt, and Keith Tyo. Metabolic kinetic modeling provides insight into complex biological questions, but hurdles remain. *Current opinion in biotechnology*, 59:24–30, October 2019.

[30] Pedro A. Saa and Lars K. Nielsen. Construction of feasible and accurate kinetic models of metabolism: A Bayesian approach. *Scientific Reports*, 6(1):29635, July 2016.

[31] Peter C. St John, Jonathan Strutz, Linda J. Broadbelt, Keith E. J. Tyo, and Yannick J. Bomble. Bayesian inference of metabolic kinetics from genome-scale multiomics data. *PLOS Computational Biology*, 15(11):e1007424, November 2019.

[32] Saratram Gopalakrishnan, Satyakam Dash, and Costas Maranas. K-FIT: An accelerated kinetic parameterization algorithm using steady-state fluxomic data. *Metabolic Engineering*, 61:197–205, September 2020.

[33] Xu Zhang, Ya Su, Andrew N. Lane, Arnold J. Stromberg, Teresa W. M. Fan, and Chi Wang. Bayesian kinetic modeling for tracer-based metabolomic data. *BMC Bioinformatics*, 24(1):108, March 2023.

[34] Subham Choudhury, Michael Moret, Pierre Salvy, Daniel Weilandt, Vassily Hatzimanikatis, and Ljubisa Miskovic. Reconstructing Kinetic Models for Dynamical Studies of Metabolism using Generative Adversarial Networks. *Nature Machine Intelligence*, 4(8):710–719, August 2022.

[35] Takashi Okada and Atsushi Mochizuki. Law of Localization in Chemical Reaction Networks. *Physical Review Letters*, 117(4):048101, July 2016.

[36] Atsuki Hishida, Takashi Okada, and Atsushi Mochizuki. Patterns of change in regulatory modules of chemical reaction systems induced by network modification. *PNAS Nexus*, page pgad441, December 2023.

[37] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[38] Yoshihiro Toya, Takanori Shiraki, and Hiroshi Shimizu. SSDesign: Computational metabolic pathway design based on flux variability using elementary flux modes. *Biotechnology and Bioengineering*, 112(4):759–768, 2015.

[39] Ingrid M. Keseler, Julio Collado-Vides, Socorro Gama-Castro, John Ingraham, Suzanne Paley, Ian T. Paulsen, Martín Peralta-Gil, and Peter D. Karp. EcoCyc: A comprehensive database resource for Escherichia coli. *Nucleic Acids Research*, 33(Database Issue):D334–D337, January 2005.

[40] Cong T. Trinh, Pornkamol Unrean, and Friedrich Srienc. Minimal Escherichia coli Cell for the Most Efficient Production of Ethanol from Hexoses and Pentoses. *Applied and Environmental Microbiology*, 74(12):3634–3643, June 2008.

[41] Bashir Sajo Mienda, Mohd Shahir Shamsir, and Rosli Md Illias. Model-guided metabolic gene knockout of gnd for enhanced succinate production in Escherichia coli from glucose and glycerol substrates. *Computational Biology and Chemistry*, 61:130–137, April 2016.

[42] Ranjini Chatterjee, Cynthia Sanville Millard, Kathleen Champion, David P. Clark, and Mark I. Donnelly. Mutation of the ptsG Gene Results in Increased Production of Succinate in Fermentation of Glucose byEscherichia coli. *Applied and Environmental Microbiology*, 67(1):148–154, January 2001.

[43] Quanfeng Liang, Fengyu Zhang, Yikui Li, Xu Zhang, Jiaojiao Li, Peng Yang, and Qingsheng Qi. Comparison of individual component deletions in a glucose-specific phosphotransferase system revealed their different applications. *Scientific Reports*, 5(1):13200, August 2015.

[44] C S Millard, Y P Chao, J C Liao, and M I Donnelly. Enhanced production of succinic acid by overexpression of phosphoenolpyruvate carboxylase in Escherichia coli. *Applied and Environmental Microbiology*, 62(5):1808–1810, May 1996.

[45] Eric W Sayers, Evan E Bolton, J Rodney Brister, Kathi Canese, Jessica Chan, Donald C Comeau, Ryan Connor, Kathryn Funk, Chris Kelly, Sunghwan Kim, Tom Madej, Aron Marchler-Bauer, Christopher Lanczycki, Stacy Lathrop, Zhiyong Lu, Francoise Thibaud-Nissen, Terence Murphy, Lon Phan, Yuri Skripchenko, Tony Tse, Jiyao Wang, Rebecca Williams, Barton W Trawick, Kim D Pruitt, and Stephen T Sherry. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 50(D1):D20–D26, December 2021.

[46] G. N. Vemuri, M. A. Eiteman, and E. Altman. Effects of Growth Mode and Pyruvate Carboxylase on Succinic Acid Production by Metabolically Engineered Strains of Escherichia coli. *Applied and Environmental Microbiology*, 68(4):1715, April 2002.

[47] Carel D. van Heerden and Willie Nicol. Continuous and batch cultures of Escherichia coli KJ134 for succinic acid fermentation: Metabolic flux distributions and production characteristics. *Microbial Cell Factories*, 12(1):80, September 2013.

[48] Roman A. Laskowski, Fabian Gerick, and Janet M. Thornton. The structural basis of allosteric regulation in proteins. *FEBS Letters*, 583(11):1692–1698, June 2009.

[49] Jean-Pierre Changeux. 50th anniversary of the word "allosteric". *Protein Science*, 20(7):1119–1124, 2011.

[50] Vetle Simensen, Christian Schulz, Emil Karlsen, Signe Bråtelund, Idun Burgos, Lilja Brekke Thorfinnsdottir, Laura García-Calvo, Per Bruheim, and Eivind Almaas. Experimental determination of Escherichia coli biomass composition for constraint-based metabolic modeling. *PLoS ONE*, 17(1):e0262450, January 2022.

[51] Xiaoping Liao, Hongwu Ma, and Yinjie J Tang. Artificial intelligence: A solution to involution of design–build–test–learn cycle. *Current Opinion in Biotechnology*, 75:102712, June 2022.