Inferring Reaction Elasticities from Metabolic Correlations in Cells Through Multi-objective Evolutionary Optimization

Arthur Lequertier[1], Wolfram Liebermeister[1], and Alberto Tonda[2,3](✉)

[1] Université Paris-Saclay, INRAE, MaIAGE, 78350 Jouy-en-Josas, France
{arthur.lequertier,wolfram.liebermeister}@inrae.fr
[2] UMR 518 MIA-PS, INRAE, Université Paris-Saclay, 91120 Palaiseau, France
alberto.tonda@inrae.fr
[3] UAR 3611 Institut des Systèmes Complexes de Paris Île-de-France (ISC-PIF), CNRS, Paris, France

Abstract. Parameter fitting in metabolic models can be challenging because experimental data are often noisy and sparse. In Bayesian estimation, prior knowledge about model parameters would be weighted against knowledge from data fitting. Since error bars and prior widths are often unknown, we explore a more flexible way of regulating this tradeoff. We propose an evolutionary multi-objective approach to parameter estimation to find compromises between parameters matching the prior (prior loss) and yielding good data fits (likelihood loss). Our metabolic model describes an ensemble of steady states with correlated variation of all model variables. In the estimation, reaction elasticities are the parameters and the covariances of measurable state variables serve as measurement data. To evaluate our approach, we conduct two tests with artificial data and a known ground truth. We first consider a simple metabolic pathway with 3 reactions and 4 metabolites, where the correlated variation of variables can be understood intuitively. The second test involves a more complex real-world metabolic model of *Escherichia coli* bacteria with 62 metabolites, 57 reactions, and 234 elasticity coefficients to be fitted, where the results are almost impossible to guess even for domain experts. In both cases, the proposed method yields satisfactory results. This paves the way to studying biological objective functions unrelated to model fitting, including homeostasis or information transmission across metabolic networks.

Keywords: Bacteria · Covariance matrix · Metabolic model · Multi-objective optimization · Parameter estimation · Structural Kinetic Model

1 Introduction

Cell metabolism consists of a network of enzyme-catalyzed chemical reactions showing a complex dynamics. Metabolic models are based on reaction networks

P. García-Sánchez et al. (Eds.): EvoApplications 2025, LNCS 15612, pp. 323–337, 2025.
https://doi.org/10.1007/978-3-031-90062-4_20

whose nodes and edges carry different types of variables – metabolite concentrations, enzyme levels, and metabolic fluxes – which depend on each other in complex ways. Physical laws, including mass balance relations and kinetic rate laws, govern the dynamics of this high-dimensional dynamical system. Experimental "omics" data, assigning values to model variables in different states of the cell, are usually scarce and noisy and often do not capture the absolute values of variables, rather measuring their relative variation between different states of the cell. This, together with limited data about model parameters, makes model parameterization a challenging task.

The Structural Kinetic Models (SKM) approach [17] is a way of formulating metabolic models that removes some of these difficulties. It describes metabolic systems in two steps, by first defining a steady reference state – a plausible set of all model variables, describing a viable state of the cell – and then modeling dynamic variations around this state with dynamics described as linear approximations and with reaction elasticities as parameters to be sampled or fitted. The reaction elasticities describe how individual reaction rates respond to changes in metabolite concentrations. A main advantage of the SKM formulation over traditional metabolic models is that it already starts from a steady state with plausible metabolic fluxes, which allows for constructing realistic models without the need for a brute-force parameter estimation.

Here we ask how the covariations of metabolic variables are shaped by network structure and details of enzyme kinetics and regulation. How much information is contained in observed correlations [16]? Focusing on covariations instead of a single steady state has different reasons: they are not only easier to measure in "omics experiments", but they also tell us more clearly how variables are dynamically related. Moreover, variance and covariances of variables may be important for cellular regulation, homeostasis, or adapted responses to changes in the cells' environment.

While covariation happens dynamically as cell variables fluctuate in time, we can also think of covariation across an ensemble of steady states, that is, states in which all variables remain constant in time, but differ between model instances, for example depending on cells' environments. To model such an ensemble of states, we may assume that some "external" variables are chosen from random distributions while all the remaining "internal" variables assume their steady-state values given those variables. The resulting variations and covariations of all variables are shaped by the structure of the metabolic network (which enforces, for example, covarying fluxes along linear metabolic pathways), but also by the reaction kinetics.

Here we study how SKM models can be fitted to covariance data. Given a network structure and a known reference state, the free parameters of an SKM are its reaction elasticities: in order to find their values, however, it is not enough to fit experimental data – in our case, the elements of its covariance matrix. It is also necessary to consider that elasticities should not differ too strongly from theoretical expectations, either because some of their values are approximately known from literature or because deviating too far from certain values might have costly side-effects for the cell. The two requirements naturally lead to a multi-objective problem, where each candidate solution will represent a trade-off between the two conflicting aims.

Below we propose a novel multi-objective evolutionary approach to parameter estimation for SKMs. In our model, we describe how covariances of model variables depend on reaction elasticities. Then we turn this question around: from a given covariance matrix of metabolite (or metabolite and enzyme) concentration data we infer the elasticities. This yields an estimate of the Jacobian matrix, an estimation problem that has been recently tackled in [10] using another methodology. In our method, the parameter estimation task is framed as an optimization problem with two conflicting aims: finding reaction elasticities that, on the one hand, fit a given covariance matrix of state variables and, on the other hand, are not too different from the theoretical expectations about approximately known or physiologically optimal parameters of the cell.

We test the approach on metabolic models following the SKM framework, first with a simple case study using a 3-reaction pathway (with 2 external and 2 internal metabolites); and then with a larger, more realistic network model describing the core metabolism of *Escherichia coli* bacteria. For simplicity, the elasticities with respect to external metabolites are assumed to be fixed and given. The results show that the proposed multi-objective optimization approach is able to find satisfying Pareto fronts for both cases. As expected, the candidate solutions match the ground truth when full data are available and deviate from the ground truth as more and more data are masked.

2 Background

This section briefly summarizes the methods used in this work: metabolic models, SKMs, and multi-objective evolutionary optimization.

2.1 Cell Metabolic Models

A metabolic model describes biochemical reactions that occur within a cell, enabling the cell to maintain its biological functions. Its nodes represent chemical species called metabolites and its edges represent the chemical reactions themselves. Here is an example of a 3-reaction linear pathway that will later serve as a toy model for the experimental evaluation:

$$A_{\text{ext}} \xrightarrow{\mathrm{R}_1} B \xrightarrow{\mathrm{R}_2} C \xrightarrow{\mathrm{R}_3} D_{\text{ext}}$$

A_{ext} and D_{ext} are external metabolites (with concentrations treated as model parameters). The concentrations of internal metabolites B and C and the reaction rates v_1, v_2, and v_3 are state variables. Each reaction follows an unknown rate law of the form $v_i = e_i \; f_i(\mathbf{c})$. The metabolites consumed and produced in each reaction are described by a stoichiometric matrix $\mathbf{N}$ whose rows represent metabolites and whose columns correspond to reactions. Each matrix element represents the stoichiometric coefficient between a metabolite and a reaction, with negative elements for reaction substrates and positive elements for reaction products. Considering the mass balance for each metabolite, we can relate the temporal variation of internal metabolite concentrations (in the vector $\mathbf{c}$) to the reaction rate $\mathbf{v}$:

$$\frac{\mathrm{d}\mathbf{c}}{\mathrm{d}t} = \mathbf{N} \cdot \mathbf{v} \Rightarrow \frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} c_B \\ c_C \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}. \tag{1}$$

We now consider a steady state in which reaction fluxes and metabolite concentrations have reference values $\mathbf{v}^*$ and $\mathbf{c}^*$, satisfying the mass conservation equation $\mathbf{N} \cdot \mathbf{v}^* = \mathbf{0}$ and the rate laws $\mathbf{v}^* = \mathbf{v}(\mathbf{e}, \mathbf{c}^*)$ To describe how small changes in metabolite concentrations influence reaction rates near this steady state, we use elasticity coefficients. The reaction elasticity E_{ij} quantifies how sensitive a reaction rate v_i is to a small change in the concentration of metabolite c_j. Based on a given rate law $v_i = v_i(e_i, \mathbf{c})$, an elasticity is defined as:

$$\mathrm{E}_{ij} = \frac{\partial v_i}{\partial \mathrm{c}_j}|_{\mathbf{c}^*, \mathbf{v}^*} \tag{2}$$

Applied to a reference state, it tells us how much the reaction rate v_i changes when c_j is slightly perturbed, assuming that all other system variables remain unchanged. Using these coefficients, we can approximate the direct effect of changes in metabolite concentrations $\delta\mathbf{c}$ on the reaction rates $\delta\mathbf{v}$:

$$\delta\mathbf{v} \approx \mathbf{E} \cdot \delta\mathbf{c}.$$

The elasticity matrix $\mathbf{E}$ contains the reaction elasticities for all the reactions and internal metabolites in the system.

2.2 Linearized Metabolic Model

In the Structural Kinetic Modeling (SKM) approach, we construct a linearized metabolic model in which unknown reaction elasticities are formally treated as model parameters [17]. The unscaled elasticity matrix can be written as

$$\mathbf{E} = \mathrm{diag}(\mathbf{v}^*)\ \boldsymbol{\mathcal{E}}\ \mathrm{diag}(\mathbf{c}^*)^{-1}$$

where $\boldsymbol{\mathcal{E}}$ is the scaled version of the elasticity matrix. The matrix $\boldsymbol{\mathcal{E}}$ is sparse: it contains non-zero entries only for metabolites directly involved in reactions. The dimensionless coefficient $\mathcal{E}_{ij}$, for the i^{th} reaction, measures the normalized degree of saturation of the catalyzing enzyme with respect to the j^{th} metabolite. It has a known sign (1 for substrates and -1 for products), and its absolute value ranges between 0 (when the enzyme is fully saturated by the metabolite) and a maximum given by the absolute stoichiometric coefficient $|n_{ji}|$ (for a fully unsaturated enzyme). We can therefore write the matrix as

$$\boldsymbol{\mathcal{E}} = -\mathbf{N}^{\top} \circ \mathbf{A} \tag{3}$$

with a matrix of saturation values a_{ij} in the range between 0 and 1. Here, $\circ$ denotes the component-wise multiplication (Hadamard product) of the two matrices. Below, for the optimization, we will represent this matrix by a vector $\mathbf{a}$ containing all the relevant (potentially non-zero) elements of the matrix. A given vector $\mathbf{a}$ of saturation values will, therefore, define a parameterized model.

To examine the dynamics of our metabolic model around a reference state, we study the Jacobian matrix, which characterizes the sensitivity of each metabolite to changes in its neighbor metabolites. The Jacobian determines how fluctuations in one metabolite can propagate across the network. For a given steady reference state, it is given by

$$\mathbf{J} = \mathbf{N}\ \mathbf{E} \tag{4}$$

where $\mathbf{N}$ is the (known) stoichiometric matrix and $\mathbf{E}$ is the unscaled elasticity matrix. By simulating the resulting metabolic dynamics, one can infer the overall behavior of the cell under different conditions and explore how the network responds to changes, despite uncertainties in the underlying kinetic details.

The SKM approach uses matrices $\mathbf{N}$ and $\mathbf{E}$, as well as fixed vectors $\mathbf{v}^*$ and $\mathbf{c}^*$, to represent the metabolism dynamic. While the stoichiometric matrix $\mathbf{N}$ represents the well-known topology of the metabolic network and plausible reference states can be guessed, the vector $\mathbf{a}$, defining the scaled elasticity matrix $\boldsymbol{\mathcal{E}}$, is typically unknown. In SKM, one may circumvent this problem by sampling this vector at random [8,12,17]; here, instead, we will fit it to data.

2.3 Correlated Variation of Metabolic Variables

To model variability or uncertainties of metabolic variables, we describe them as random variables. Given random distributions of the external variables (external metabolite concentrations and enzyme levels), we obtain distributions of all the state variables (internal metabolite concentrations and fluxes). The model variables are described on a logarithmic scale and their variations are assumed to be small, allowing us to describe the system's dynamics in a linear approximation, using notions from Metabolic Control Analysis (MCA) [9,14]. Our aim is to compute a global linear response of all internal variables of the system (internal metabolite concentrations $\mathbf{c}$ and reaction fluxes $\mathbf{v}$) to external variables (enzyme levels and external metabolite concentrations) that serve as the sources of perturbations. This global linear response is captured by the unscaled response matrix [9]

$$\mathbf{R_p} = \begin{pmatrix} \mathbf{R_p^c} \\ \mathbf{R_p^v} \end{pmatrix} \quad \text{where} \quad \begin{aligned} \mathbf{R_p^c} &= -\mathbf{L}{\mathbf{J}_{\mathrm{R}}}^{-1}\mathbf{N}_{\mathrm{R}} \cdot \mathbf{E_p} \\ \mathbf{R_p^v} &= \mathbf{E} \cdot \mathbf{R_p^c} + \mathbf{E_p}. \end{aligned}$$

In the formula., the reduced stoichiometric matrix $\mathbf{N}_{\mathrm{R}}$ consists of a set of linearly independent rows of $\mathbf{N}$, and the link matrix is defined to obtain $\mathbf{N}$ again via $\mathbf{N} = \mathbf{L}\mathbf{N}_{\mathrm{R}}$ [14]. The reduced Jacobian is given by $\mathbf{J}_{\mathrm{R}} = \mathbf{N}_{\mathrm{R}}\ \mathbf{E}\ \mathbf{L}$.

From now on we assume that all variables are described on a logarithmic scale. In analogy to the scaled elasticities, we define a scaled version $\boldsymbol{\mathcal{R}}_{\mathbf{p}}$ of the response matrix $\mathbf{R_p}$, which is used in this case. Moreover, rather than considering specific perturbations, we consider random perturbation and treat all variables as random variables [7,11,18]. On a logarithmic scale, all variables are assumed to follow normal distributions around the given reference state. The (logarithmic) external variables (in a random vector $\mathbf{P}$) are assumed to be independent, with a predefined, diagonal covariance matrix $\mathrm{Cov}(\mathbf{P})$. Due to the linearized model, also the resulting (logarithmic) state variables will follow normal distributions. Altogether, we obtain a random vector $\mathbf{Z}$ comprising all the (logarithmic) external and internal variables, following a joint multivariate normal distribution with covariance matrix [11]

$$\mathrm{Cov}(\mathbf{Z}) = \begin{pmatrix} \mathbf{I_p} \\ \boldsymbol{\mathcal{R}}_{\mathbf{p}} \end{pmatrix} \cdot \mathrm{Cov}(\mathbf{P}) \cdot \begin{pmatrix} \mathbf{I_p} \\ \boldsymbol{\mathcal{R}}_{\mathbf{p}} \end{pmatrix}^{\mathrm{T}} \tag{5}$$

where $\mathbf{I_p}$ is the identity matrix with a size equal to the number of external variables in $\mathbf{p}$. Equation (5) is a compact representation of how uncertainty in the network's external environment leads to uncertainty in all variables. The covariance matrix $\mathrm{Cov}(\mathbf{Z})$ depends on the covariance matrix $\mathrm{Cov}(\mathbf{P})$ of external variables and on the scaled response matrix $\boldsymbol{\mathcal{R}}_{\mathbf{p}}$, which itself depends on the stoichiometric matrix $\mathbf{N}$ and the vector of saturation levels $\mathbf{a}$.

2.4 Inferring Reaction Elasticities from Covariances in Model Variables

For a model with given reference state, stoichiometric matrix, and random distributions of the external variables, our aim is to estimate the unknown saturation levels in $\mathbf{a}$ based on data by assuming two different objectives. The first objective is to fit experimental data by minimizing the *negative* log likelihood between the model output and experimental data – in our specific case, a covariance matrix derived from experimental data in different steady states. Since in practice, data are limited, we assume that only some of the covariances are available for the estimation. The second objective reflects our prior knowledge about plausible saturation values in $\mathbf{a}$, described by a prior distribution. For the prior, we assume a Gaussian distribution with given mean vector and standard deviations. The mean values represent our best guess for biologically reasonable levels of saturation. This second objective, effectively, penalizes deviations of the saturation values from their prior means.

Thus, our two-objective approach balances two key considerations: (1) the likelihood that the model accurately reflects experimental data, and (2) the prior knowledge or assumptions about plausible saturation values based on established biochemical principles. These objectives can be in conflict, as measurement data may suggest values that deviate significantly from prior expectations. This trade-off between empirical alignment and adherence to prior knowledge represents a central challenge in metabolic modeling.

2.5 Multi-objective Evolutionary Optimization

Machine learning and optimization algorithms have been applied to metabolic network models [2]. Among these methods, Evolutionary Algorithms (EAs) have emerged as a powerful class of optimization techniques [1]. These approaches also allow for multi-objective optimization and that the cell does not have a single stable state but rather a set of preferred modes, each optimizing, to a greater or lesser extent, key characteristics of homeostasis.

When dealing with multiple conflicting objectives, classic optimization techniques might employ a composite objective function defined as a weighted sum of the single objective functions. The idea behind multi-objective optimization [4], instead, is to not assign weights to the functions at all but to explore the space of trade-offs between the conflicting objectives. The result is not just one single best solution, but a set of candidate solutions that are not Pareto-dominated by others. Being population-based, EAs are well suited to multi-objective optimization and currently represent the state of the art in the domain [15]. In this work, we chose to employ the established Non-Sorting Genetic Algorithm II (NSGA-II) [5]: while not particularly recent, it is still extremely competitive in optimization problems with up to 3 objectives and has implementations in multiple programming languages, from C++ to Python.

3 An Approach for Model Fitting by Multi-objective Optimization

Given the two conflicting objectives of parameter estimation in metabolic models described above, we propose a novel approach based on multi-objective evolutionary optimization to find a set of good compromise solutions to be later analyzed for their biological relevance. A scheme of our estimation problem with the two objective functions is presented in Fig. 1.

3.1 Structure of a Candidate Solution

In our optimization problem, an individual represents a set of reaction elasticities, that capture the local sensitivities of reaction fluxes to changes in metabolite concentrations. The elasticity matrix is encoded by a saturation level vector $\mathbf{a}$, whose elements are real-valued numbers in $[0, 1]$. An individual $\mathbf{a}$ thus represents a configuration of the SKM model from which we compute the model's covariance

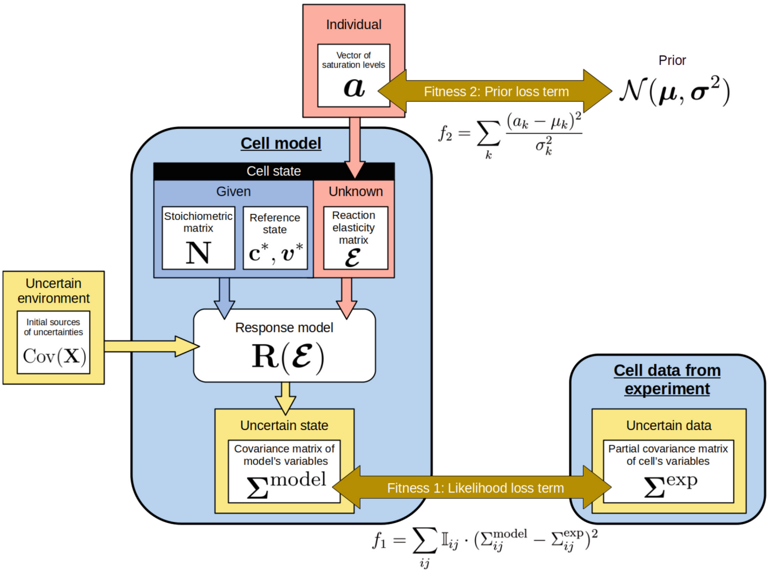
Individual
Vector of saturation levels
$\boldsymbol{a}$
Fitness 2: Prior loss term
Prior
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$
$f_2 = \sum_k \frac{(a_k - \mu_k)^2}{\sigma_k^2}$
Cell model
Cell state
Given
Unknown
Stoichiometric matrix
$\mathbf{N}$
Reference state
$\mathbf{c}^*, \boldsymbol{v}^*$
Reaction elasticity matrix
$\boldsymbol{\mathcal{E}}$
Uncertain environment
Initial sources of uncertainties
$\mathrm{Cov}(\mathbf{X})$
Response model
$\mathbf{R}(\boldsymbol{\mathcal{E}})$
Uncertain state
Covariance matrix of model's variables
$\boldsymbol{\Sigma}^{\mathrm{model}}$
Fitness 1: Likelihood loss term
Cell data from experiment
Uncertain data
Partial covariance matrix of cell's variables
$\boldsymbol{\Sigma}^{\mathrm{exp}}$
$f_1 = \sum_{ij} \mathbb{I}_{ij} \cdot (\Sigma_{ij}^{\mathrm{model}} - \Sigma_{ij}^{\mathrm{exp}})^2$

Fig. 1. Estimation problem for metabolic models. The aim is to estimate reaction elasticities, using covariances of state variables as data. A model instance is defined by a stoichiometric matrix $\mathbf{N}$, a known reference state ($\mathbf{v}^*$ and $\mathbf{c}^*$), and unknown reaction elasticities $\mathcal{E}_{ij}$, parameterized by saturation values in a vector $\mathbf{a}$. A model instance yields an input-output relationship described by a response coefficient matrix $\mathbf{R}$. Based on an assumed random distribution of external variables ("environment"), the model generates a covariance matrix of all the model variables. In our estimation procedure for saturation values, we consider two objectives. The first loss function f_1 compares the covariance results between our model and experimental data. The second loss function f_2 compares the elasticity matrix to a prior. Solutions are found by an evolutionary algorithm in which each individual represents a vector of saturation values, encoding an instance of our response model.

matrix, $\mathrm{Cov}(\mathbf{Z}^{\mathrm{model}})$. During our estimation procedure, this covariance matrix is then scored by one of our two objective functions.

Given the structure of a candidate solution – a simple numerical vector – the operators that will be used during the evolutionary optimization are a 1-point crossover and a polynomial mutation [6].

3.2 Objective Functions

Our first objective concerns the similarity between a covariance matrix $\mathbf{\Sigma}^{\mathrm{model}} = \mathrm{Cov}(\mathbf{Z}^{\mathrm{model}})$ obtained from a model and an "experimentally measured" covariance matrix $\mathbf{\Sigma}^{\mathrm{exp}} = \mathrm{Cov}(\mathbf{Z}^{\mathrm{exp}})$. The likelihood function (assuming a normal distribution for "measurement errors" of covariance values with a constant standard deviation σ), is given by

$$\mathcal{L}(\mathbf{a}^{\text{model}}|\boldsymbol{\Sigma}^{\text{exp}}) = \prod_{ij} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\Sigma_{ij}^{\text{model}}(\mathbf{a}^{\text{model}}) - \Sigma_{ij}^{\text{exp}})^2}{2\sigma^2}\right).$$

In our optimization, we consider the negative log-likelihood and ignore constant terms. Our resulting first objective function f_1, called "likelihood loss", can be written as follows:

$$f_1(\mathbf{a}^{\text{model}}) = \sum_{ij} \mathbb{I}_{ij} \cdot (\Sigma_{ij}^{\text{model}}(\mathbf{a}^{\text{model}}) - \Sigma_{ij}^{\text{exp}})^2 \tag{6}$$

with $\mathbb{I}_{ij} \in \{0, 1\}$: 1 if a data point is accessible, 0 otherwise, reflecting the impossibility of observing specific matrix elements from biological data. A maximal likelihood corresponds to a minimal loss.

For the second objective, we score each possible vector $\mathbf{a}$ by a prior density $\mathrm{P}_{\text{prior}}(\mathbf{a})$, an uncorrelated multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and standard deviations in a vector $\boldsymbol{\sigma}$. With this prior, a saturation value vector $\mathbf{a}$ can be scored by how much it deviates from the prior mean. In analogy to our loss function f_1, we define our second objective function f_2, called "prior loss":

$$f_2(\mathbf{a}^{\text{model}}) = \sum_{k} \frac{(a_k^{\text{model}} - \mu_k)^2}{\sigma_k^2}. \tag{7}$$

The function represents the negative log-prior, where constant terms are again ignored.

4 Experimental Evaluation

To validate our approach, we ran computer experiments on two different models. We first considered a simple 3-reaction pathway for which the results are easy to analyze. Then we considered the *E. coli* core model [13], a standard network model of *Escherichia coli* bacteria consisting of 62 metabolites and 57 reactions, and with a total of 234 reaction elasticities to be estimated.

In both cases, the artificial data used in the estimation were generated using a "true" instance of our model, taken to be our ground truth. The true saturation values were chosen randomly within biologically plausible ranges. In the models to be fitted, we kept all model parameters exactly the same except for the saturation values in $\mathbf{a}$ to be estimated. The prior mean values for all saturation values were set to 1/2, describing a case in which enzymes are half-saturated with all the metabolites. Biochemically, this corresponds to metabolite concentrations matching their respective Michaelis-Mention constants K_{M}. All prior standard deviations were chosen to be equal, and also all data error bars (for covariance values) were chosen to be equal: the two numerical values do not play a role,

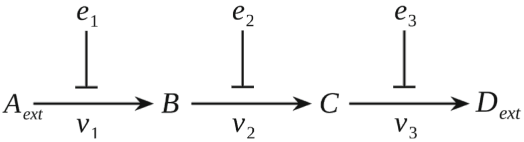

Fig. 2. Metabolic pathway of 3 reactions. Variability in external variables (enzyme levels and concentrations of external metabolites $\mathrm{A_{ext}}$ and $\mathrm{D_{ext}}$) causes variability in internal variables (reaction fluxes and concentrations of internal metabolites B and C).

because their only effect is a linear scaling of the two objective functions, which will not change the shape of our Pareto front.

All the necessary code for reproducing the experiments is available in a public GitHub repository[1]; the scripts are implemented in Python 3, resorting to the `pymoo` library [3] for NSGA-II.

4.1 Simple 3-Reaction Pathway Model

As a first test for of our optimization algorithm, we studied a hypothetical 3-reaction linear pathway (Fig. 2). The internal variables are the fluxes v_1, v_2 and v_3 and the concentrations c_{B} and c_{C}, while the external variables (and therefore sources of uncertainty) are the external metabolite concentrations $c_{A_{\mathrm{ext}}}$ and $c_{D_{\mathrm{ext}}}$, as well as the enzyme activities e_1, e_2 and e_3. The elasticity matrix $\boldsymbol{\mathcal{E}}$ (for internal metabolites B and C) contains only two columns, and only 4 of its elements are non-zero. Therefore, a candidate solution is represented by a saturation vector of length 4, $\mathbf{a}^{model} \in [0,1]^4$, parameterizing the reaction elasticities with respect to internal metabolites. This minimal setup enables a rapid and easily interpretable analysis, which allowed us to identify and understand challenges during the implementation.

After a few trial runs, we ran NSGA-II with the following hyperparameters: population size $\mu_e = 1000$, offspring size $\lambda_e = 1000$, tournament selection with $\tau = 2$, probability of crossover $p_c = 0.9$, probability of polynomial mutation [6] $p_m = 0.9$, and a stop condition after $G_{max} = 100$ generations. The experiment took about 20 min to run on a server with 72 Intel Xeon w9-3475X CPUs and 128 GB of RAM, with one evaluation taking around 0.02 s on average.

The results are shown in Fig. 3 (left). The plots show how the algorithm progressively identifies high-quality, non-dominated points in each iteration, ultimately converging toward a satisfactory front. Since the objective functions are positive with optimal (minimal) values of 0, the fact that the front almost touches both axes indicates a good performance. The two red dots at the ends of the front represent the known extreme points; although they are shown here for reference, they were not used in computing the front.

[1] https://github.com/albertotonda/evolutionary-optimization-cell-models.

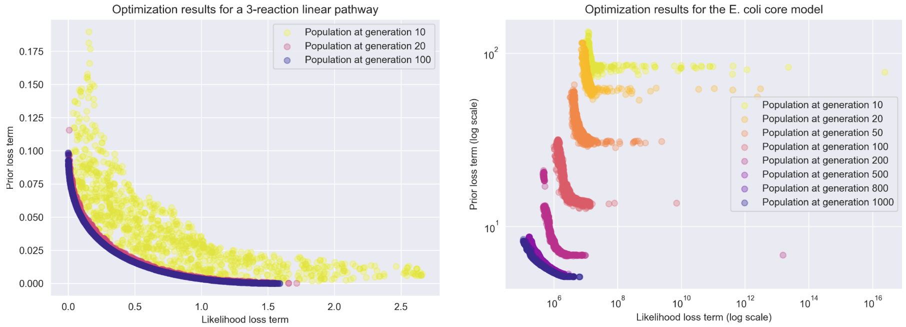

Fig. 3. Population of NSGA-II at selected generations for the 3-reactions pathway model (see Fig. 2) and the *E. coli* core model. Both objectives (prior and likelihood loss term) are minimized. Left: Experiment on the 3-reaction pathway. Right: Experiment on the *E. coli* core model. (Color figure online)

Our objective f_1 compares a covariance matrix produced by a candidate solution to our "true" the covariance matrix, replacing here covariances from biological data. In real biological experiments, only some of the cellular variables can be measured. We studied the effects of such incomplete data by considering a number of scenarios in which only some of the covariance data were used to compute the likelihood loss f_1. Figure 4 shows optimization results for different scenarios with such incomplete data. To indicate the varying quality of the estimation, the distance of each parameter set from the ground truth (the **a** vector of the "true model") is shown in color. The "true model" itself is represented by the red dot on the top left. The red point on the bottom right represents an individual whose values match the prior mean.

In Scenario 1 (top left), the whole covariance matrix is considered (a 10×10 matrix – but the symmetric shape of the matrix allows us to reduce it to 55 elements). Our Pareto front connects these two points. Distances between the individuals and the "true" saturation values are coherent compared with the values of the first objective function along the Pareto front. The individual that is the closest to the top-left red point has the lowest loss value, which suggests the uniqueness of the solution in this case. In Scenario 2 (top right), only the covariances of metabolites and enzymes (but not the fluxes) are considered, reducing to 13 the number of elements compared. Distances from the ground truth remain well sorted along the Pareto front. With more iterations, the front would probably extend to the two red points. In Scenario 3 (bottom left), only covariances between metabolites are considered, reducing to 7 the number of elements compared. The distances from the ground truth (in color) remain globally coherent, but do not fully match the first objective function. This means that an estimation based on metabolite covariances still works, but not fully reliably. In Scenario 4 (bottom right), only the covariance between internal metabolites was considered, reducing to 3 the number of elements compared. As expected,

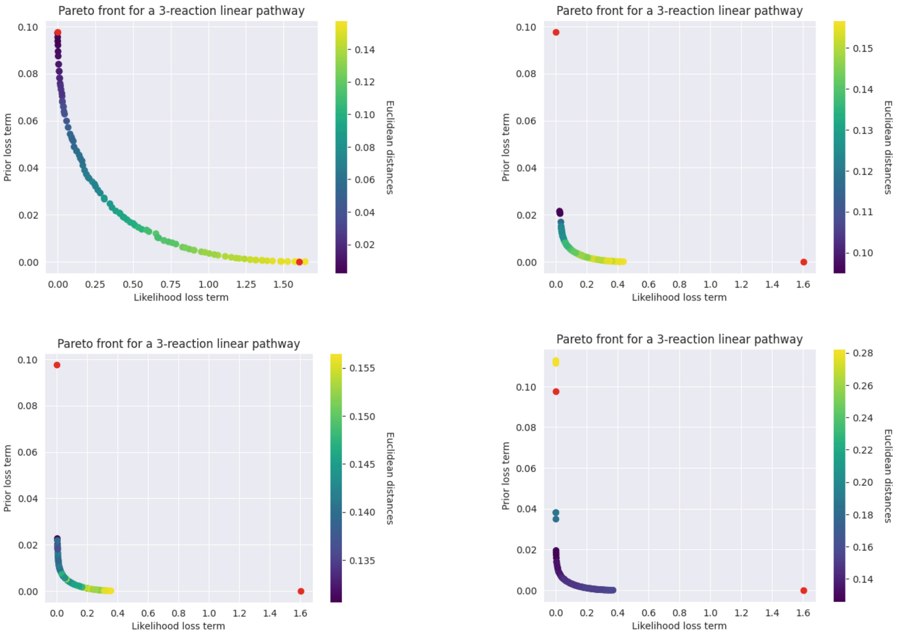

Fig. 4. Population of NSGA-II (last generation) after optimizing the 3-reaction pathway using different estimation scenarios. Top left (Scenario 1): Loss function f_1 based on covariances between all model variables (enzyme levels, metabolite concentrations, and fluxes). Top right (Scenario 2): f_1 based on covariances of 3 enzyme levels and 4 metabolite concentrations. Bottom left (Scenario 3): f_1 based on covariances of 4 metabolite concentrations. Bottom right (Scenario 4): f_1 based on covariances of 2 internal metabolite concentrations. (Color figure online)

with 3 data points and 4 parameters to be estimated, the estimation problem is ill-determined and the estimates along the entire front are partially shaped by the prior.

4.2 *Escherichia Coli* core model

Our second computer experiment targets the *E. coli* core model [13], which contains 62 metabolites, 57 reactions, and 234 considered reaction elasticities. A candidate solution is now represented by a vector **model** $\in [0,1]^{234}$, again representing enzyme saturation values encoding reaction elasticities. As the problem is now more complex, a larger computational budget was allocated to NSGA-II, with hyperparameters: $\mu_e = 1000, \lambda_e = 1000, \tau = 2, p_c = 0.9, p_m = 0.9$, and $G_{max} = 1000$ generations. The experiment took about 96 h to run on a server with 72 Intel Xeon w9-3475X CPUs and 128 GB of RAM, with one evaluation taking on average 0.34 s.

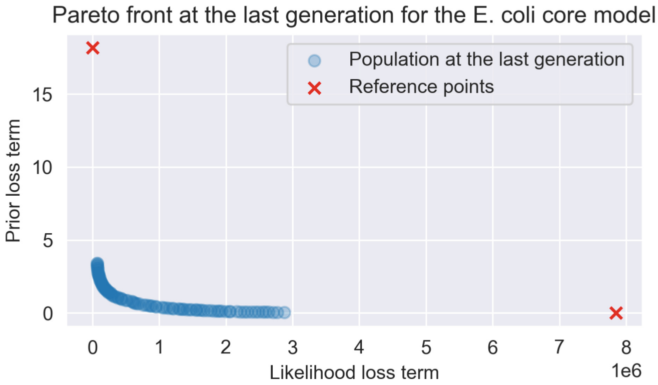

Fig. 5. Pareto front at generation 1000 for the *E. coli* core model.

The results from the evolutionary run on the *E. coli* core model are presented in Fig. 3 (right). Figure 5 shows the set of non-dominated points found at the last generation, along with the two reference points for the known optimal solutions. While the point at the bottom right represents the known prior mean, the point at the top left represents the "true" vector **a** behind the artificial data, which would not be accessible if the ground truth were not known. The results are satisfying: like with the smaller test model, the Pareto front comes very close to the optimal values of the single objectives.

5 Conclusions and Future Works

To explore how measured covariances in metabolite and enzyme data could be used for parameter estimation, we proposed a framework for multi-objective optimization in metabolic models. In our models, a know random distribution of external metabolite and enzyme concentrations leads to a simple formula for covariances between all model variables. We now used a given covariance matrix to infer some model parameters (the enzyme saturation values with respect to internal metabolites), assuming all other model parameters to be known.

In Bayesian estimation, a parameter set would be scored by its posterior density, a product of likelihood and prior. Since the relative importance of these terms may vary (depending on prior widths and data error bars), we treated them here as separate objectives. Each point on the Pareto front represents a model instance. Since the assumed "true" parameter set (our ground truth) differs from the prior mean, there is a trade-off between the objectives, resulting in an extended front that we recovered in our computer experiments. Points from the two ends of the front represent, respectively, solutions that are well supported by data (but with the risk of overfitting) versus more conservative solutions that stay close to our prior expectation about model parameters. By moving along the front, we can interpolate between these extremes and shift our focus of attention between prior and likelihood. This would not be possible in a standard Bayesian estimation.

Each point of the front corresponds to a parametrization of the Jacobian matrix which determines the metabolic system dynamics. Biochemically, the inferred elasticities are important: they can tell us whether reactions are forward-driven (insensitive to their product concentrations, which also makes them relatively enzyme-efficient) or, in contrast, close to chemical equilibrium (sensitive to substrate and product concentrations alike, which may imply a high enzyme demand). In our model, thermodynamic driving forces may also be included explicitly by using the Structural Thermokinetic Modeling variant of SKM [12].

In our optimization, we assumed that reference state and elasticities for external metabolites were known, and estimated only the elasticities for internal metabolites. This could be generalized to fit other model details, including reference fluxes or concentrations, thermodynamic forces, or the presence of regulatory arrows. Likewise, instead of covariances also other types of data could be considered, such as time series or sets of different steady states. With only minor modifications, the proposed framework can also be used to study biological objectives, for example, trade-offs between dynamic robustness and enzyme costs.

Acknowledgments. The authors would like to acknowledge funding from the ANR project Artificial Metabolic Networks (ANR-21-CE45-0021), a grant from the French National Agency for Research.

References

1. Ananda, R., Daud, K.M., Zainudin, S.: Non-dominated sorting differential search algorithm for optimizing regulatory-metabolic networks by using probabilistic approach. In: 2023 International Conference on Electrical Engineering and Informatics (ICEEI), pp. 1–6. IEEE (2023). https://doi.org/10.1109/iceei59426.2023.10346837
2. Bai, L., et al.: Advances and applications of machine learning and intelligent optimization algorithms in genome-scale metabolic network models. Syst. Microbiol. Biomanufacturing **3**(2), 193–206 (2022). https://doi.org/10.1007/s43393-022-00115-6
3. Blank, J., Deb, K.: pymoo: multi-objective optimization in python. IEEE Access **8**, 89497–89509 (2020)
4. Deb, K.: Multi-Objective Optimization Using Evolutionary Algorithms, vol. 16. John Wiley & Sons (2001)
5. Deb, K., Pratap, A.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**, 182–197 (2002)
6. Deb, K., Sindhya, K., Okabe, T.: Self-adaptive simulated binary crossover for real-parameter optimization. In: GECCO '07, Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation, pp. 1187–1194. Association for Computing Machinery, New York, NY, USA (2007)
7. Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S.: Stochastic gene expression in a single cell. Science **297**(5584), 1183–1186 (2002). https://doi.org/10.1126/science.1070919

8. Grimbs, S., Selbig, J., Bulik, S., Holzhütter, H.G., Steuer, R.: The stability and robustness of metabolic states: identifying stabilizing sites in metabolic networks. Mol. Syst. Biol. **3**, 146 (2007)
9. Hofmeyr, J.: Metabolic control analysis in a nutshell. In: ICSB 2001 Online Proceedings. http://www.icsb2001.org/toc.html (2001)
10. Li, J., Weckwerth, W., Waldherr, S.: Network structure and fluctuation data improve inference of metabolic interaction strengths with the inverse jacobian. npj Syst. Biol. Appl. 137 (2024)
11. Liebermeister, W., Klipp, E.: Biochemical networks with uncertain parameters. IEE Proc. Sys. Biol. **152**(3), 97–107 (2005)
12. Liebermeister, W.: Structural thermokinetic modelling. Metabolites **12**(5), 434 (2022). https://doi.org/10.3390/metabo12050434
13. Orth, J.D., Fleming, R.M.T., Palsson, B.O.: Reconstruction and use of microbial metabolic networks: the core escherichia coli metabolic model as an educational guide. EcoSal Plus **4**(1) (2010). https://doi.org/10.1128/ecosalplus.10.2.1
14. Reder, C.: Metabolic control theory: a structural approach. J. Theor. Biol. **135**(2), 175–201 (1988). https://doi.org/10.1016/s0022-5193(88)80073-0
15. Sharma, S., Kumar, V.: A comprehensive review on multi-objective optimization techniques: past, present and future. Arch. Comput. Methods Eng. **29**(7), 5605–5633 (2022). https://doi.org/10.1007/s11831-022-09778-9
16. Steuer, R., Kurths, J., Fiehn, O., Weckwerth, W.: Observing and interpreting correlations in metabolomic networks. Bioinformatics **19**(8), 1019–1026 (2003). https://doi.org/10.1093/bioinformatics/btg120
17. Steuer, R., Gross, T., Selbig, J., Blasius, B.: Structural kinetic modeling of metabolic networks. Proc. Natl. Acad. Sci. **103**(32), 11868–11873 (2006)
18. Uhlendorf, J., Bockmayr, A., Liebermeister, W.: Prediction of optimal enzymatic regulation architectures. Master's thesis, Department of Mathematics and Computer Science Bioinformatics Program (2009)