Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 58-14 (2024) 422–427

Elucidation of Macroscopic Stoichiometry and Kinetics of Bioprocesses using Sparse Identification

Guilherme A. Pimentel * **Fernando N. Santos-Navarro** *
Laurent Dewasme * **Alain Vande Wouwer** *
* *Systems, Estimation, Control and Optimization (SECO) ,*
University of Mons, 7000 Mons, Belgium
(e-mail: <guilherme.araujopimentel, fernandonobel.santosnavarro,
laurent.dewasme, alain.vandewouwer>@umons.ac.be).

Abstract: This paper presents a systematic data-driven methodology to infer macroscopic reaction schemes and their associated kinetic laws from the measurements of concentration trajectories. The procedure uses sparse identification incorporated with a generic kinetic structure combining activation and inhibition factors. Only measurements of the extracellular species, i.e., biomass, substrates, and products of interest, are required, and measurement noise can be tackled using specific regularization techniques. The methodology is illustrated with a case study of a synthetic dataset from the production of therapeutic proteins using mammalian cell cultures.

Keywords: mathematical modeling, sparse identification, stoichiometry, kinetics, parameter estimation, biotechnology

1. INTRODUCTION

Macroscopic modeling of bioprocesses is particularly important in developing software sensors and model-based controllers. This task typically implies two fundamental steps: (i) the definition of a reaction scheme and the estimation of the stoichiometry, and (ii) the inference of a kinetic structure. Both problems have aroused considerable interest in the literature (Grosfils et al., 2007b).

On the one hand, a candidate macroscopic reaction scheme can be determined by applying the Principal Component Analysis (PCA) to the time evolution of the measurements of the macroscopic species under consideration (Bernard and Bastin, 2005). An extension of this method, called Maximum Likelihood Principal Component Analysis (MLPCA), was proposed by Mailier et al. (2012a) to account for higher levels of measurement noise. More recently, Pimentel et al. (2023b) exploited a data-driven methodology to deduce macroscopic reaction schemes using a robust algorithm for parallel implicit sparse identification proposed by Kaheman et al. (2020).

On the other hand, the problem of determining the kinetic laws has been mainly tackled from two different angles, either combining several known kinetic laws such as the one of Monod (Monod, 1949) for substrate activation or Jerusalimski (Jerusalimski and Engamberdiev, 1969) for inhibition, or through the proposal of more generic and polyvalent laws either inspired by existing kinetic model structures (Haag et al., 2003; Grosfils et al., 2007a) or by black box approaches such as neural networks (Vande Wouwer et al., 2004). For instance, Mailier and Vande Wouwer (2012b) developed a likelihood ratio test to choose the most likely kinetic structure among candidate models. The latter method involves examining a large number of model alternatives, which unfortunately grows combinatorially and makes it unsuitable for large-scale metabolic networks. The same issue affects the approaches proposed by Mangan et al. (2016) and Kaheman et al. (2020). However, Wang et al. (2020) proposed an alternative approach, assuming that the combination of Monod and Jerusalimski factors covers all possible kinetic modulation effects of the involved compounds. In the same spirit, Grosfils et al. (2007a); Richelle and Bogaerts (2015) proposed a simple but powerful method for extracting the activation and inhibition factors involved in the reaction rates for a predefined reaction scheme. This method consists of obtaining the reaction rates signal from the process derivative measurements and using the logarithm transformation to rewrite the problem into a linear identification problem that reveals the process compound involved in the activation and inhibition of the reaction rates. Recently, Forster et al. (2023) proposed a methodology very close to the one of Grosfils et al. (2007a); Richelle and Bogaerts (2015) but with a significantly more complex optimization based on mixed-integer nonlinear programming (MINLP). Dewasme et al. (2023) also proposed a practical data-driven modeling procedure combining MLPCA and the systematic selection of Monod/Jerusalimski laws based on local parametric sensitivities.

This study aims to extend the results of Pimentel et al. (2023b) to include the inference of kinetic laws to the sparse identification procedure. One of the basic features of sparse identification is the selection of a dictionary of basis functions that could adequately describe the system's nonlinearity. Here, we make use of the general kinetic structure proposed by Richelle and Bogaerts (2015), which is amenable to a linearization through a logarithmic transformation. This greatly simplifies the extraction of information about activation and inhibition mechanisms. In a further step, the kinetics can be expressed as products of simple Monod/Jerusalimski laws in order to conform to the tradition in bioprocess modeling. The resulting method dramatically streamlines the size of the candidate library of functions for parameter identification and yields a systematic and efficient methodology to extract biological models out of experimental data.

10.1016/j.ifacol.2024.08.373

This paper is organized as follows. Section 2 details the macroscopic modeling approach, and Section 3 presents the data-driven techniques involved in the proposed method. Section 4 shows the results and analysis of a case study of the protein production by mammalian cells considering lactate shift. Section 5 presents the conclusion and future work.

2. BIOPROCESS MACROSCOPIC MODELING

A macroscopic reaction scheme is a set of M reactions involving N key species, which are typically biomass, substrates, and products (Bastin and Dochain, 1990):

$$\sum_{i\in\mathcal{R}_j} k_{i,j}\xi_i \xrightarrow{\varphi_j(\xi,\vartheta_j)} \sum_{l\in\mathcal{P}_j} k_{l,j}\xi_l \tag{1}$$

where $\mathcal{R}_j$ and $\mathcal{P}_j$ denote the sets of reactants, represented by ξ_i, and products, defined as ξ_l, in the j^{th} reaction. $k_{i,j}$ and $k_{l,j}$ are pseudo-stoichiometric coefficients while $\varphi_j(\xi,\vartheta_j)$ are the corresponding reaction rates, functions of ξ (reactant/product quantities or concentrations) and the parameters of the rate kinetic structure ϑ_j.

Applying mass balance to (1), the following ordinary differential equation system is obtained:

$$\frac{d\xi(t)}{dt} = K\varphi(\xi(t),\vartheta) + v(\xi(t),t), \tag{2}$$

where K is the pseudo-stoichiometric matrix, and $v(\xi(t),t)$ represents the transport term, including dilution effects, input feeds, and gaseous outflows. In most cases, the number of components N is larger than the number of reactions M so that the rank of the stoichiometric matrix K is assumed to be M. In an identification study, the component concentrations $\xi(t)$ are measured and the transport terms $v(\xi(t),t)$ are known/measurable so that a so-called transport-free mass-balance system can be expressed (Mailier et al., 2012a) as

$$\frac{d\xi^\star(t)}{dt} = K\varphi(\xi^\star(t),\vartheta), \tag{3}$$

where $\dot{\xi}^\star(t) = \dot{\xi}(t) - v(\xi(t),t)$ is the transport-free state vector, and $\dot{\xi}^\star(t)$ denotes the time derivative of $\xi^\star(t)$.

3. DATA-DRIVEN METHODS

Reaction rates are intrinsically driven by a few reactants and products, inducing activation and inhibition effects in a sparse combination. This property makes the sparse identification framework an adequate approach.

3.1 Parallel and Implicit Sparse Identification

Consider the following general nonlinear dynamic system

$$\frac{d\xi(t)}{dt} = f(\xi(t)), \tag{4}$$

where $\xi(t)$ is the state vector $\xi(t) = [\xi_1(t)\ \cdots\ \xi_N(t)]^T \in \mathbb{R}^N$, and the system dynamics is represented by a function $f(\xi(t))$, which could be described in terms of a library of functions

$$\Theta(\xi) = [\theta_{lib,1}(\xi)\ \theta_{lib,2}(\xi)\ \cdots\ \theta_{lib,w}(\xi)], \tag{5}$$

where w is the number of elements. Thus, each row equation may be written as

$$\frac{d\xi_k(t)}{dt} = f_k(\xi(t)) \approx \Theta(\xi)\Omega_k, \tag{6}$$

where Ω_k is a sparse vector, indicating which terms are active in the dynamics (Brunton et al., 2016).

To determine the nonzero entries of Ω_k through sparse regression based on the trajectory data, the time-series data is arranged into a matrix $\Xi = [\xi(t_1),\xi(t_2)\ \cdots\ \xi(t_{n_s})]^T$, and the associated derivative matrix $\dot{\Xi} = \left[\dot{\xi}(t_1),\dot{\xi}(t_2)\ \cdots\ \dot{\xi}(t_{n_s})\right]^T$ is computed using appropriate numerical differentiation methods.

It is now possible to describe the dynamical system using a model that is linear in the parameters and evaluated with the measured state trajectories:

$$\dot{\Xi} = \Theta(\Xi)\Omega. \tag{7}$$

Equation (7) might also involve derivatives of the state variables on the right-hand side, i.e., include a factor $\Theta(\Xi,\dot{\Xi})$ and to solve implicit model structures, Kaheman et al. (2020) proposed a constrained optimization formulation where each candidate function is tested individually in an implicit and parallel optimization. However, each of these individual equations may be combined into a single constrained system of equations

$$\Theta(\Xi,\dot{\Xi}) = \Theta(\Xi,\dot{\Xi})\Omega \quad \text{such that } \Omega_{jj} = 0. \tag{8}$$

The constraint $\Omega_{jj} = 0$ forces the solution not to be the trivial one ($\Omega = I_{w\times w}$) and the optimization problem can be written as

$$\begin{aligned} &\min_{\Omega} \|\Theta(\Xi,\dot{\Xi}) - \Theta(\Xi,\dot{\Xi})\Omega\|_2, \\ &\text{s.t. } diag(\Omega) = 0, \text{and } \forall|\Omega_{\{i,j\}}| < \lambda, \Omega_{\{i,j\}} = 0, \end{aligned} \tag{9}$$

where λ is a sparsity-promoting parameter. This problem can be solved in various ways, but in this study, the sparsity pattern is obtained using sequentially thresholded least squares, which iteratively computes a least-squares solution to minimize (9). Any element of Ω smaller than a threshold λ is set to zero, and then (9) is solved again with these fixed zero elements. The sparsity parameter λ is a hyper-parameter, and each column equation may require a different parameter λ_y (Kaheman et al., 2020). In particular, to solve problem (9), CVX is used, a package for specifying and solving convex programs (Grant and Boyd, 2014).

The procedure is simple and consists of organizing the measurements and their derivatives in the vector $\Theta(\Xi,\dot{\Xi})$, while a large value is given to λ. Then, the value of this parameter is decreased, and the fitting error $||\dot{\Xi} - \hat{\dot{\Xi}}||_2/||\dot{\Xi}||_2$ is analyzed for each of the identified state derivatives $\hat{\dot{\Xi}}$, for instance. A model candidate is obtained when the error is small, and the vector Ω is sparse. This procedure is repeated for each different library Θ.

3.2 Computing Derivatives

Computing the measurement derivatives can be challenging due to measurement scarcity and noise. The scientific literature proposes various numerical differentiation methods, such as filter-based approaches, Tikhonov regularization, and smoothing splines, which have been successful in different applications (Varah, 1982). Unfortunately, the mathematical formulation of numerical differentiation is typically ill-posed, and one often resorts to an *ad hoc* selection of one of the numerous computational methods.

In this study, Butterworth filtering is combined with the method of van Breugel et al. (2020) to estimate the time derivative of the measurement signals. The approach of van Breugel et al. (2020) is a multi-objective optimization framework where a set of parameters can be fine-tuned to estimate the derivatives of noisy data.

3.3 Selection of the Dictionary of Kinetic Laws (Activation, Saturation, and Inhibition Factors)

Our approach will consider successively two distinct libraries of basis functions to represent the nonlinear kinetic rates. The first library is intended to discover activation and inhibition mechanisms, and the second library is intended to get a model in the popular form of Monod/Jerusalimski factors.

First, the reaction rates are represented by the kinetic model structure proposed in (Grosfils et al., 2007a; Richelle and Bogaerts, 2015):

$$\varphi_j(\xi) = \alpha_j \prod_{m\in R_i} \xi_m^{\gamma_{i,m}} \prod_{l\in P_i} e^{-\beta_{i,l}\xi_l} \tag{10}$$

where $\alpha_j > 0$ is the j^{th} rate constant, $\gamma_{i,m} \geq 0$ the activation coefficient of component m in the j^{th} reaction, $\beta_{j,l} \geq 0$ the inhibition coefficient of component l in the j^{th} reaction. R_j and P_j are, respectively, the sets of indices of the reactants and products that activate and/or inhibit reaction j. Note that Eq. (10) is only able to account for a simple saturation effect by the combination of activation and inhibition factors of the same compound (Grosfils et al., 2007b; Richelle and Bogaerts, 2015). However, the main advantage of (10) is the possibility to linearize the kinetic model structure concerning the parameters by a logarithmic transformation of the form:

$$\ln\varphi_j(\xi,t) = \ln\alpha_j + \textstyle\sum_h \gamma_{hj} \ln\xi_h(t) - \sum_l \beta_{lj}\xi_l(t), \quad j\in[1,M], \tag{11}$$

under the constraints:

$$[\gamma_{h\ldots j} \ \ \beta_{l\ldots j}] \geq 0, \tag{12}$$

which results in a sparse vector of γ and β factors, easily inferred by a linear identification procedure (in the original papers, a simple least-square solution is used).

Therefore, to cast the framework of the general kinetic model structure into the parallel and implicit sparse identification (Section 3.1), we first relax the constraint $\Omega_{jj} = 0$, replacing it by $\Omega_{11} = 0$. This eases the implementation of the constraints imposed on γ and β (see (12)). From Eq. (11), $\varphi_j(\xi,t)$, which is obtained by knowing the process stoichiometry and the numerical differentiation of a predefined compound measurement, can be expressed by the combination of the reactants and the products involved in the process. Thus, the following general library is used for each reaction:

$$\Theta_j(\hat{\Xi},\hat{\Xi}) = [\ln\varphi_j(\xi,t) \ \ 1 \ \ \ln\xi_1(t) \ \cdots \ \ln\xi_h(t) \ \ -\xi_1(t) \ \cdots \ -\xi_l(t)], \tag{13}$$

$$\Omega_j = [0 \ \ \alpha_j \ \ \gamma_{1j} \ \cdots \ \gamma_{hj} \ \ \beta_{1j} \ \cdots \ \beta_{lj}]^T, \tag{14}$$

where $j = [1,\cdots,M]$.

Once the activation and inhibition effects have been unveiled by the values of $\gamma_{hj} \neq 0$ and $\beta_{lj} \neq 0$, a second library of basis functions can be proposed in terms of classical Monod and Jerusalimski factors:

$$\varphi_j(\xi) = \mu_{max,j} \prod_{m\in R_j} \frac{\xi_m}{\xi_m + K_{j,\xi m}} \prod_{l\in P_j} \frac{K_{j,I\xi}}{K_{j,I\xi l} + \xi_l} X, \tag{15}$$

i.e., in a well-accepted form in the study of biological models, where $\mu_{max,j}$, is the maximum specific rate of reaction j, K_{j,ξ_m} and $K_{j,I\xi l}$ are respectively the saturation and inhibition constants of reaction j. In *Remark 1*, we present a simple example of how to build this library, and *Remark 2*, we highlight the interest of sparse identification to achieve this last step.

Remark 1. Identifying Monod law structures and their parameters using the direct application of the sparse identification framework is unfeasible in cases where one wants to reveal if a desired compound activates or inhibits the reaction rates. For instance, let us consider a simple case of activation/inhibition kinetics, i.e.,

$$\mu(P) = \mu_{max,P} \frac{P}{P+K_P} \frac{K_{I,P}}{P+K_{I,P}}, \tag{16}$$

where K_P and $K_{I,P}$ are the half-saturation and inhibition constants, respectively. Following the general sparse identification procedure, (16) *can be developed as follows:*

$$(K_P K_{I,P})\cdot\mu(P) + (K_P + K_{I,P})\cdot\mu(P)P + 1\cdot\mu(P)P^2 = (\mu_{max,P}K_{I,P})\cdot P \tag{17}$$

and a regressive form where the regressor $\Theta(\Xi)$ and the parameter vector Ω read:

$$\Theta(\Xi) = \begin{bmatrix}\mu(P) & \mu(P)P & \mu(P)P^2 & -P\end{bmatrix}, \tag{18}$$

$$\Omega = [(K_P K_{I,P}) \quad (K_P + K_{I,P}) \quad 1 \quad (\mu_{max,P}K_{I,P})]. \tag{19}$$

Consider now that P is a measured activating compound, and $\mu(P)$ can be deduced from the derivative of the measurements of P. Sparse identification should reveal that $K_{I,P} = 0$ and $K_P \neq 0$. However, canceling $K_{I,P}$ in Ω implies that the problem is structurally unidentifiable as it is impossible to infer neither the kinetic structure nor the parameter values. This illustrates the importance of an a priori identification of the compound influences on the kinetics, which also allows a dramatic decrease in the number of candidate functions of the library vector $\Theta(\Xi)$.

Remark 2. The parameters of the Monod and Jerusalimski factors could be estimated using a classical nonlinear least squares (NLS) approach based on the results of the previous sparse identification which discovers the activating/inhibiting compounds in each reaction. One could, therefore, wonder why we propose an additional sparse identification step at this stage of the procedure. The main reason is that sparse identification does not require parameter initialization whereas a classical NLS method is sensitive to initialization when the optimization problem is multimodal, i.e., possesses several local minima. Sparse identification, therefore, provides a good initial estimate of the kinetic parameters for a further NLS identification step (see next Section 3.4). Hence, parameter initialization is achieved in a data-driven manner.

3.4 Global Identification

The separated identification of the stoichiometry and kinetics may result in a parameter estimation bias. To solve this issue, a global adjustment of the full parameter set can be achieved using the previously identified stoichiometric and kinetic parameter values as initial guesses in a conventional nonlinear identification procedure. A numerical optimizer from MATLAB can be used for this purpose ("fmincon", "fminsearch", "lsqnonlin" or a combination of the latter), minimizing the distance between the vector of observations ξ_η and macroscopic model predictions $\xi(\theta)$ in a nonlinear least-squares cost function of the form:

$$J(\theta) = \sum_{w=1}^{N_{exp}} \sum_{i=1}^{n_s} (\xi_{i,w}(\theta) - \xi_{\eta,i,w})^T Q_{i,w}^{-1} (\xi_{i,w}(\theta) - \xi_{\eta,i,w}), \tag{20}$$

where θ is the vector of parameters to be identified, index i denotes the sample time of the w^{th} experiment, and $Q_{i,w}$ is a diagonal scaling matrix with the squares of the maximum concentration levels (this allows scaling with respect physical units and magnitudes). Parametric sensitivities can also be computed by integration of the following ordinary differential equation $\dot{\xi}_{\theta_k,j} = \frac{\partial f_j}{\partial \xi_j}\frac{\partial \xi_j(i)}{\partial \theta_k} + \frac{\partial f_j}{\partial \theta_k}$, where $\xi_{\theta_k,j}(i)$ is the sensitivity of the j^{th} state ξ_j with respect to the k^{th} parameter θ_k at time i,

with $\dot{\xi}_j = f_j$ the model RHS in (2). Parameter identifiability can be assessed using the Fisher Information Matrix (FIM), computed as:

$$FIM = \varepsilon^{-2} \sum_{i=1}^{n_s} \xi_{\theta,i}^T Q_{i,w}^{-1} \xi_{\theta,i}, \tag{21}$$

where ε^2 is the a posteriori estimate of the relative measurement error variance, inferred from the cost function residual $\varepsilon^2 = \frac{J}{n_s N_{exp} - p}$ (where p is the number of parameters). An optimistic estimate (i.e., a lower bound) of the parameter estimation error covariance matrix is obtained from the inverse of the FIM:

$$Cov(\theta) > FIM^{-1} \tag{22}$$

4. CASE STUDY: PROTEIN PRODUCTION CONSIDERING LACTATE SHIFT

Basically, mammalian cells consume glucose, its main carbon source, and produce lactate. The accumulation of lactate leads to cell growth and protein production inhibition. Cells are also likely to shift their metabolism in case of a low glycolysis rate (i.e., glucose consumption rate) combined with the presence of a significant lactate concentration. This shift triggers lactate consumption to compensate for the insufficient amount of glucose. A three-reaction model has been developed in Pimentel et al. (2023a):

$$k_{31} G + Gn \xrightarrow{\varphi_1} k_{11} X_v + k_{51} L + k_{61} P \tag{23a}$$

$$k_{52} L \xrightarrow{\varphi_2} X_v \tag{23b}$$

$$X_v \xrightarrow{\varphi_3} X_d + k_{63} P \tag{23c}$$

where X_v, X_d, G, Gn, L, and P are, respectively, the concentrations of viable biomass, dead biomass, glucose, glutamine, lactate, and proteins. The first reaction involves the consumption of glucose and glutamine to produce viable biomass, lactate, and proteins. This reaction is regulated by φ_1. The second reaction consists of the consumption of lactate to generate viable biomass, governed by φ_2. Finally, the third reaction represents the death of viable biomass, leading to the production of dead biomass and the release of proteins into the medium. Applying mass balance to (23) yields the following ordinary differential equation system:

$$\frac{dX_v}{dt} = k_{11}\varphi_1 + \varphi_2 - \varphi_3, \tag{24a}$$

$$\frac{dX_d}{dt} = \varphi_3, \tag{24b}$$

$$\frac{dG}{dt} = -k_{31}\varphi_1, \tag{24c}$$

$$\frac{dGn}{dt} = -\varphi_1, \tag{24d}$$

$$\frac{dL}{dt} = k_{51}\varphi_1 - k_{52}\varphi_2, \tag{24e}$$

$$\frac{dP}{dt} = k_{61}\varphi_1 + k_{63}\varphi_3, \tag{24f}$$

where the reaction rates are defined as:

$$\varphi_1 = \mu_{max,1} \frac{Gn}{(K_{Gn} + Gn)} \frac{G}{(K_G + G)} X_v, \tag{25a}$$

$$\varphi_2 = \mu_{max,2} \frac{L}{(K_L + L)} \frac{K_{GI}}{(K_{GI} + G)} X_v, \tag{25b}$$

$$\varphi_3 = \mu_{dmax} \frac{K_{Gnd}}{(K_{Gnd} + Gn)} X_v, \tag{25c}$$

K_{Gn}, K_G, K_L are the half-saturation parameters, $\mu_{max,1}$, $\mu_{max,2}$, and μ_{dmax} the maximum reaction rate parameters, and K_{GI} and K_{Gnd} the inhibition parameters. The reaction rate φ_1 is driven by two Monod factors activated by glucose and glutamine. Likewise, φ_2 stands for the selective consumption of lactate activated by lactate and inhibited by the presence of glucose. φ_3 models the biomass death rate inhibited by the presence of glutamine, which is the primary nitrogen source of the cell, ensuring its viability.

Simulations of a cell batch culture were performed considering independent and identically distributed (IID) Gaussian noise $e \sim (0, \sigma^2)$ corrupting the measurements. The imposed variances are 0.1×10^6cell/ml, 0.0167×10^6cell/ml, 0.2 g/l, 0.01 g/l, 0.1 g/l, and 1.0 μg/ml for viable biomass X_v, dead biomass X_d, glucose G, glutamine Gn, lactate L, and proteins P, respectively, which are taken four times a day ($t_s = 0.25$ *days*) for a culture time of seven days ($t_{batch} = 7$ *days*).

4.1 Step 1: Number of Reactions and Stoichiometry

This step of the whole methodology is the subject of an earlier publication (Pimentel et al., 2023b) and is not described in detail in this paper. Here, let us focus on the case study and say that the data samples are first processed using the derivative approximation method presented in Section 3.2, delivering $(\hat{\Xi}, \dot{\hat{\Xi}})$. These numerical derivatives are used as library functions, i.e.,

$$\Theta(\hat{\Xi}, \dot{\hat{\Xi}}) = [\hat{X}_v \quad \hat{X}_d \quad \hat{G} \quad \hat{G}_n \quad \hat{L} \quad \hat{P}]. \tag{26}$$

Then, sparse identification is applied, and the following results are obtained with the corresponding values of the sparsity parameter λ:

$$\hat{G} = \quad 17.858\hat{G}_n \qquad \lambda = 10 \tag{27a}$$

$$\hat{P} = \quad 4.1955\hat{X}_d - 102.76\hat{G}_n \qquad \lambda = 2 \tag{27b}$$

$$\hat{L} = -1.1379\hat{X}_v - 1.2632\hat{X}_d - 18.124\hat{G}_n \qquad \lambda = 1 \tag{27c}$$

$$\hat{X}_v = -1.0699\hat{X}_d - 15.984\hat{G}_n - 0.88484\hat{L} \qquad \lambda = 0.5 \tag{27d}$$

$$\hat{X}_d = -0.21364\hat{X}_v - 2.3115\hat{G}_n - 0.21863\hat{L} \qquad \lambda = 0.08 \tag{27e}$$

$$\hat{G}_n = -0.064311\hat{X}_v - 0.055947\hat{X}_d - 0.052474\hat{L} \qquad \lambda = 0.02 \tag{27f}$$

The inferred number of reactions is the maximum number of species concentration derivatives involved in each equation of (27), i.e., three. Combining the relations found in (27) with three assumptions based on some process *a priori* knowledge, the values of the stoichiometric parameters of the macroscopic reactions can be calculated. The first assumption is that a growth reaction is associated with glucose and glutamine. Thus, we can define a first reaction rate $\rho_1 = -\hat{\hat{G}}_n$ (we use another notation to make a possible distinction with the rates φ_j of the original model) and use (27a) to obtain $\hat{\hat{G}} = -17.858\rho_1$, which delivers k_{31}, see (24c). The second assumption considers a death reaction, where we define $\rho_2 = \hat{\hat{X}}_d$ (obviously $\rho_1 = \varphi_1$, but $\rho_2 = \varphi_3$). Using the definitions of ρ_1 and ρ_2 combined with (27b), the stoichiometric parameters k_{61} and k_{63} can be found (see (24f)). Note that the remaining equations, (27c) to (27f), convey the same information. Thus, using ρ_1 and ρ_2 with (27c) yields $\hat{k}_{52}$ and a second term that is the relation $\hat{k}_{51} = 18.0643 - 1.1301\hat{k}_{11}$. To obtain the values of $\hat{k}_{51}$ and $\hat{k}_{11}$, a third assumption is required. Thus a new library $\Theta(\Xi, \dot{\Xi}) = [\hat{X}_v \quad \hat{X}_d \quad \hat{G} \quad \hat{\hat{G}}_n \quad \hat{P}]$ is defined, which no longer considers $\hat{L}$. This results in the new relation $\hat{X}_v = 1.5296\hat{X}_d - 7.3401\hat{\hat{G}}_n$ from where we can obtain $\hat{k}_{11}$ and in turn $\hat{k}_{51}$ using the previous relation. Table 1 presents the identified parameters in the *Step 1 Ident* column.

Table 1. Identified Parameters of model (24). Step 1 is the identification of the stoichiometric parameters. Step 2 is the identification of the kinetic parameters. The global identification considers the full parameter set. σ_{err} is their respective relative standard deviations of estimation errors.

Parameter	Original	Step 1 Ident	Step 2 Ident	σ_{err}	Global Ident	σ_{err}
$\mu_{max,1}$	0.460	—	0.33629	2.246	0.36234	0.46351
$\mu_{max,2}$	0.400	—	0.45528	1.4241	0.30113	0.17892
μ_{dmax}	0.03	—	0.014408	0.53245	1.8859	39.834
K_G	1.10	—	2.5748	34.716	3.9616	1.1699
K_{Gn}	0.250	—	0.1805	34.323	0.044876	1.0061
K_L	1.20	—	1.2668	1.5872	0.85507	0.35308
K_{GI}	0.800	—	1.083	2.9575	1.1219	0.78463
K_{gnd}	0.002	—	0.0026574	0.92427	2.7585e-05	39.944
k_{11}	6.80	7.3401	—	0.28706	6.3768	0.088159
k_{31}	18.0	17.858	—	0.018053	18.733	0.0058337
k_{51}	10.70	9.7689	—	0.22384	11.709	0.05156
k_{52}	1.20	1.1301	—	0.1016	1.1888	0.028271
k_{61}	107.80	102.76	—	0.034894	108.99	0.010779
k_{63}	2.90	4.1955	—	1.507	2.4735	0.37776

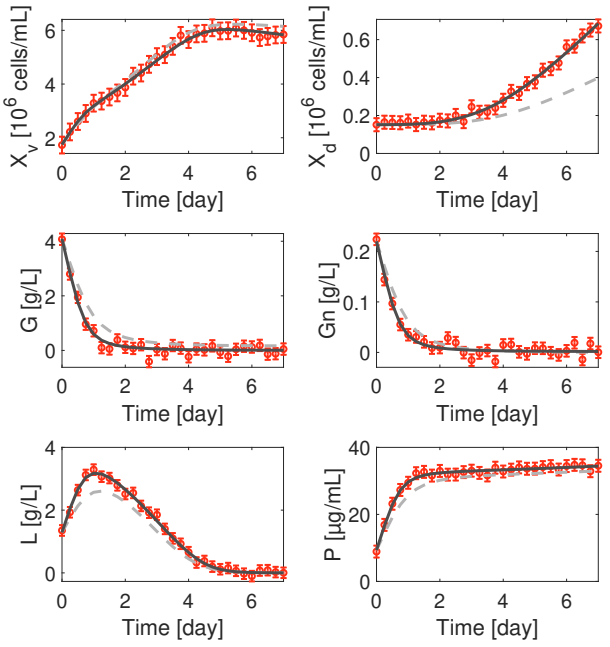

Fig. 1. Red error bars are the noisy process measurements, light-gray dashed lines are the sparse identification considering Step 1+2, and dark-gray lines show the results of the global identification.

Table 2. Activation and Inhibition terms for the reaction rates

Parameter	$\mu_{1,gen}(\cdot)$	$\mu_{2,gen}(\cdot)$	$\mu_{3,gen}(\cdot)$
α	-1.2939	-0.7868	-4.3143
γ_G	0.1368	0	0
γ_{G_n}	0.8309	0	0
γ_L	0	1.2768	0
β_G	0	0.3037	0
β_{G_n}	0	0	8.9093
β_L	0	0.8502	0
λ	0.1	0.01	0.9

4.2 Step 2: Revealing Kinetic Structures and Parameters

Two distinct libraries of basis functions are used to estimate the nonlinear kinetic rates. The first intends to discover the activation and inhibition mechanisms, and the second uses previous results to define a model in the popular form of Monod/Jerusalimski factors.

Therefore, in the first step, we use the general library (13), which in this case reads:

$$\Theta(\hat{\Xi},\dot{\hat{\Xi}}) = \left[ln\left(\hat{\mu}_i(\cdot)\right)\ \mathbf{1}\ ln(\hat{G})\ ln(\hat{G}_n)\ ln(\hat{L})\ -\hat{G}\ -\hat{G}_n\ -\hat{L}\right], \quad (28)$$

where $i = [1,\cdots,M]$. $\hat{\mu}_i(\cdot) = \hat{\varphi}_i/\hat{X}_v$ is obtained from the measurement derivatives divided by $\hat{X}_v$. Indeed, the rates are respectively approximated by $\hat{\varphi}_1 = -\hat{\dot{G}}_n$, $\hat{\varphi}_2 = (\hat{k}_{51}/\hat{k}_{52})\hat{\dot{G}}_n - \hat{\dot{L}}/\hat{k}_{52}$, and $\hat{\varphi}_3 = \hat{\dot{X}}_d$ (for simplicity of analysis of the results we use back the notation of the reference model). Then, using (9) and sweeping λ to minimize the errors of each of the estimates, we obtain the values presented in Table 2.

The second step to estimate the parameters of nonlinear kinetic rates in the form of Monod/Jerusalimski factors exploits the results of this latter Table. From the second column, the structure of the first reaction rate can be defined as $\hat{\mu}_1(G,G_n)$, activated by glucose and glutamine Monod factors imposed by $\gamma_G \neq 0$ and $\gamma_{G_n} \neq 0$. Accordingly, the suggested library and the unknown parameter vector are composed of the reorganization of the Monod terms as a sum of its factors, which reads:

$$\Theta(\Xi,\dot{\Xi}) = \left[\hat{\mu}_1\hat{G}_n\hat{G} \quad \hat{\mu}_1\hat{G}_n \quad \hat{\mu}_1\hat{G} \quad \hat{\mu}_1 \quad \hat{G}_n\hat{G}\right], \quad (29)$$

$$\Omega = \left[\mathbf{1} \quad K_G \quad K_{Gn} \quad K_GK_{Gn} \quad -\mu_{max,1}\right]^T. \quad (30)$$

In the same way, the third column of Table 2 shows that lactate drives both activation and saturation ($\gamma_L \neq 0$ and $\beta_L \neq 0$) while glucose is an inhibitor ($\beta_G \neq 0$). The corresponding regression vectors read:

$$\Theta(\Xi,\dot{\Xi}) = \left[\hat{\mu}_2 \quad \hat{\mu}_1\hat{G} \quad \hat{\mu}_2\hat{L} \quad \hat{\mu}_2\hat{L}\hat{G} \quad \hat{L}\right], \quad (31)$$

$$\Omega = \left[K_LK_{Gi} \quad K_L \quad K_{Gi} \quad 1 \quad -\mu_{max,2}K_{Gi}\right]^T. \quad (32)$$

The last column of Table 2 shows that $\hat{\mu}_3$ is inhibited by glutamine and Θ becomes:

$$\Theta(\Xi,\dot{\Xi}) = \left[\hat{\mu}_3\hat{G}_n \quad \hat{\mu}_3 \quad 1\right], \quad (33)$$

$$\Omega = \left[1 \quad K_{gnd} \quad -\mu_{dmax}K_{gnd}\right]^T. \quad (34)$$

Table 1 reports the identified parameter values. The *Step 2 Ident* column shows fair values compared to the reference ones, while the relative standard deviations of the estimation errors remain in an acceptable range. It is important to highlight that all the data-driven procedures have been implemented in a decentralized way. The measurement derivative estimates are considered independently, and the identification of the kinetic parameters is run by considering each rate separately. For instance, φ_1 affects $\dot{X}_v$, $\dot{G}$, $\dot{G}_n$, $\dot{L}$, and $\dot{P}$. Still, it has been

inferred assuming $\hat{\varphi}_1 = -\hat{G}_n$. Therefore, a global estimation of the full model is required to cancel the remaining estimation bias.

4.3 Global Identification

The noisy measurements and the proposed model predictions are respectively contained in the vectors ξ_η and $\xi(\theta)$. In addition, we consider the full parameter set vector

$$\theta = [\mu_{max,1}\ \mu_{max,2}\ K_{Gn}\ K_G K_L\ K_{GI}\ \mu_{dmax}\ K_{gnd}\ k_{11}\ k_{31}\ k_{51}\ k_{52}\ k_{61}\ k_{63}\ X_{v0}\ X_{d0}\ G_0\ G_{n,0}\ L_0\ P_0] \quad (35)$$

which also includes the initial concentrations as parameters. The parameter adjustments and the relative standard deviation of the estimation errors are computed according to (20) and (22). Figure 1 highlights the fitting improvement, while the last two columns of Table 1 report the new relative standard deviation values. It is worth noting that the values of μ_{dmax} and K_{gnd}, in Step 2, are related to the range of values of the regressor $\hat{\mu}_3$ in (33), which is a result of the product $\mu_{dmax}K_{gnd}$ in (34). This can also be noticed in the global identification where μ_{dmax} is two orders of magnitude larger than the original value. In comparison, K_{gnd} is two orders of magnitude lower, and both parameters exhibit a large relative standard deviation. This shows that only the product of these parameters is practically identifiable.

5. CONCLUSION

This paper presents a unified data-driven approach based on the sparse identification framework to infer macroscopic reaction schemes, the corresponding stoichiometry, and the kinetic structures. A mammalian cell culture mechanistic model is proposed as a case study validating the proposed method. Future work entails improving the data-driven approach for accurately differentiating measurement trajectories since they are sensitive to process noise. Another perspective is the use of data-driven methods to estimate reaction rates, removing the need for numerical differentiation.

ACKNOWLEDGMENT

The authors acknowledge the support of the ProtoDrive project (convention *no. 2010119*) of the Win2Wal program of the Walloon Region (DGO6) achieved in collaboration with the CER Groupe and Univercells Exothera. The scientific responsibility rests with its authors.

REFERENCES

Bastin, G. and Dochain, D. (1990). *On-Line Estimation and Adaptive Control of Bioreactors*. Volume 1 of Process Measurement and Control, Elsevier: Amsterdam.

Bernard, O. and Bastin, G. (2005). On the estimation of the pseudo-stoichiometric matrix for macroscopic mass balance modelling of biotechnological processes. *Mathematical Biosciences*, 193(1), 51–77.

Brunton, S.L., Proctor, J.L., and Kutz, J.N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15), 3932–3937.

Dewasme, L., Mäkinen, M., and Chotteau, V. (2023). Practical data-driven modeling and robust predictive control of mammalian cell fed-batch process. *Computers and Chemical Engineering*, 171(108164), 1–16.

Forster, T., Vázquez, D., Cruz-Bournazou, M.N., Butté, A., and Guillén-Gosálbez, G. (2023). Modeling of bioprocesses via minlp-based symbolic regression of s-system formalisms. *Computers & Chemical Engineering*, 170, 108108.

Grant, M. and Boyd, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1.

Grosfils, A., Vande Wouwer, A., and Bogaerts, P. (2007a). On a general model structure for macroscopic biological reaction rates. *Journal of biotechnology*, 130(3), 253–264.

Grosfils, A., Vande Wouwer, A., and Bogaerts, P. (2007b). Systematic decoupled identification of pseudo-stoichiometry, degradation rates and kinetics. *Computers & chemical engineering*, 31(11), 1449–1455.

Haag, J.E., Vande Wouwer, A., and Remy, M. (2003). A general model of reaction kinetics in biological systems. In *2003 European Control Conference (ECC)*, 2929–2934. IEEE.

Jerusalimski, N. and Engamberdiev, N. (1969). *Continuous cultivation of microorganisms*, volume 517. Academic Press, New York.

Kaheman, K., Kutz, J.N., and Brunton, S.L. (2020). SINDy-PI: a robust algorithm for parallel implicit sparse identification of nonlinear dynamics. *Proceedings of the Royal Society A*, 476 (2242).

Mailier, J. and Vande Wouwer, A. (2012b). Identification of nested biological kinetic models using likelihood ratio tests. *Chemical Engineering Science*, 84, 727–734.

Mailier, J., Remy, M., and Vande Wouwer, A. (2012a). Stoichiometric identification with maximum likelihood principal component analysis. *Journal of Mathematical Biology*, 67(4), 739–765.

Mangan, N.M., Brunton, S.L., Proctor, J.L., and Kutz, J.N. (2016). Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2(1), 52–63.

Monod, J. (1949). The growth of bacterial cultures. *Annual Review of Microbiology*, 3(1), 371–394.

Pimentel, G.A., Dewasme, L., Santos-Navarro, F.N., Boes, A., Côte, F., Filée, P., and Vande Wouwer, A. (2023a). Macroscopic dynamic modeling of metabolic shift to lactate consumption of mammalian cell batch cultures. In *9th International Conference on Control, Decision and Information Technologies*.

Pimentel, G.A., Dewasme, L., and Vande Wouwer, A. (2023b). On the number of reactions and stoichiometry of bioprocess macroscopic models: an implicit sparse identification approach. In *22nd World Congress of the International Federation of Automatic Control*.

Richelle, A. and Bogaerts, P. (2015). Systematic methodology for bioprocess model identification based on generalized kinetic functions. *Biochemical Engineering Journal*, 100, 41–49.

van Breugel, F., Kutz, J.N., and Brunton, B.W. (2020). Numerical differentiation of noisy data: A unifying multi-objective optimization framework. *IEEE Access*, 8, 196865–196877.

Vande Wouwer, A., Renotte, C., and Bogaerts, P. (2004). Biological reaction modeling using radial basis function networks. *Computers & chemical engineering*, 28(11), 2157–2164.

Varah, J.M. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing*, 3(1), 28–46.

Wang, M., Risuleo, R.S., Jacobsen, E.W., Chotteau, V., and Hjalmarsson, H. (2020). Identification of nonlinear kinetics of macroscopic bio-reactions using multilinear Gaussian processes. *Computers & Chemical Engineering*, 133, 106671.