Practical data-driven modeling and robust predictive control of mammalian cell fed-batch process

L. Dewasme[a], M. Mäkinen[b], V. Chotteau[b]

[a]*Systems, Estimation, Control and Optimization (SECO) Group, Faculty of Engineering, University of Mons, 31, Boulevard Dolez , 7000 Mons, Belgium (e-mail: Laurent.Dewasme@umons.ac.be).*
[b]*Centre for Advanced Bioproduction, KTH Royal Institute of Technology, Sweden (e-mail: Veronique.Chotteau@biotech.kth.se).*

Abstract

Even if the performances of bioprocesses can be significantly improved by model-based control, there often remains a tradeoff between model complexity and control robustness. This paper proposes an original data-driven strategy for fast design of dynamic bioprocess models with minimal complexity (i.e., minimal number of bioreactions). Maximum likelihood principal component analysis (MLPCA) is applied to infer the minimal reaction scheme from a 25-state mammalian cell culture database. Then, a systematic algorithm is used to provide a continuous kinetic model formulation assuming all rates to occur simultaneously, which may be far from true cell metabolic conditions sometimes presenting discontinuous metabolic switches. A robust model predictive formulation is therefore adopted to reduce the impact of model structural uncertainty on the process performances. Additional numerical results show that the proposed strategy presents excellent performances in presence of unexpected metabolic switches.

Keywords: Data-driven modeling; Animal cell culture; Fed-batch process; Model predictive control

1. Introduction

The potential of Process Analytical Technologies (PAT) in therapeutic product manufacturing becomes more and more important following the high demand for monoclonal antibodies (MAbs), viral vectors, recombinant proteins, etc. The potential benefits of data science and predictive modeling and control strategies in the fields of therapeutics are twofold: (i) by reducing the required experimental

sessions in time and money thanks to the predictive capacity of the mathematical models constituting the process digital twins, and (ii) by providing model structure properties well-adapted to the control framework.

Mechanistic modeling is by far the most plebiscited technique for animal cell cultures, from the seminal work of [1], where a one-reaction model is proposed, to more complex formulations such as in [2] or [3] where overflow metabolism (or short-term Crabtree effect [4]) as enunciated by [5], is used to describe a double bottleneck in the assimilation of glucose and glutamine, considered as main substrates. Other microscopic and accurate models have been proposed, considering metabolic network analysis and elementary flux modes as in [6, 7, 8, 9, 10, 11, 12, 13]. However, even if these strategies rely on the a priori knowledge extracted from cell reduced and, sometimes, coarse metabolic networks, conferring good predictive capabilities, they require a sufficient amount of available informative data from the cells, clearly exhibiting each of the considered metabolic phenomena (i.e., substrate overflows, cell decay, alternative substrate consumptions when starving, amino acid metabolism, etc). Data-driven modeling, limiting the model structure to the observable phenomena from the data, therefore appears as a promising alternative solution in view of accelerating the parameter identification procedure for a specific process with reproducible properties (which is often the case for bioprocesses operated in GMP conditions).

While data-driven structures may often be assimilated to machine learning or more general black-box representations ([14],[15],[16]), [17] have recently proposed a systematic method based on a maximum likelihood principal component analysis (MLPCA), inspired from the works of [18] and [19], resulting in a macroscopic mechanistic-like model formulation of hybridoma cell cultures with a minimal number of pseudo-reactions, in some sense lumping the numerous occurring network reactions and, in turn, a minimal number of stoichiometric parameters. Moreover, the mechanistic formulation based on first principles also allows to confer more predictive capabilities to the model in opposition to black-box models which often require an important amount of data to provide equivalent result performances ([20]). Moreover, conversely to hybrid modeling where certain unknown dynamics of the model are replaced by black-box structures such as artificial neural networks [21, 22], MLPCA generates a data subspace basis which can be converted into a mechanistic model formulation. The main motivation is therefore to combine the mentioned advantages of a data-driven method with first-principle knowledge through a two-step procedure aiming at (i) setting the model complexity (the minimal number of macro-reactions) based on the data informative content and (ii) transforming the resulting sub-space basis into a consistent

stoichiometric basis. However, the kinetic structure definition is not provided by the MLPCA. In a recent work, [23] have proposed a systematic method based on the multilinear gaussian process (MGP) framework, to identify key kinetic factors describing activation/saturation (i.e., the Monod law) and inhibition effects ([24]), based on the a priori knowledge of the reaction network and its stoichiometry. The first part of the current work is dedicated to the combination of an original kinetic design method inspired from [23] and MLPCA, in order to first set a data-driven reaction network of the animal cell culture fed-batch process under study and then, to identify the kinetic factors and related parameters.

The estimation of the model dimension (i.e., number of reactions) provided by the MLPCA does not take into account possible metabolic switches inducing some discontinuous reaction activation/deactivation (as it may be the case for instance in the bottleneck assumption of [5]). This modeling uncertainty issue can however be tackled when designing fed-batch process feedback control using a robust framework definition avoiding or, at least, limiting inaccurate tracking of optimal operating conditions as proposed by [25], [26], [27], [28], [29] and [30]. While only plant/model parameter mismatch is solved in the latter, an extension to structural mismatch has recently been proposed by [31] in a nonlinear model predictive control (NMPC) policy based on a set of structurally different reduced-network models. A multi-stage NMPC is then applied, optimizing the fed-batch process input trajectory with respect to several possible scenarios implying the available models. In this spirit, the second part of this paper is dedicated to the application of a multi-stage NMPC policy considering, conversely to [31] using reduced networks presenting multiple reactions and, therefore, a higher complexity level, a sole and data-driven model with minimal complexity and possible rate activation/deactivation scenarios.

The main motivations of this work are then (i) to systematically and quickly generate data-driven models of bioprocesses with large number of state variables, successively providing the minimal reaction scheme and the related kinetics and (ii) to limit the impact of model structural uncertainty in the presence of metabolic switches by formulating a robust NMPC strategy. This paper is organized as follows. Section 2 presents the data-driven MLPCA and kinetic generation strategies as well as an original parameter set reduction algorithm. Section 3 briefly describes the experimental materials and methods as well as the available data set content, and presents the direct and cross-validations of the proposed strategy, discriminating among two potential models. Section 4 is dedicated to the NMPC policy formulation while section 5 shows and comments the validation of the model as well as the comparative results between the applications of the pro-

posed multi-stage NMPC versus the classical economic NMPC. Conclusions are drawn in section 6.

2. An original data-driven modeling strategy

2.1. STEP 1: reaction scheme and stoichiometry generations by MLPCA

Classical bioprocess macroscopic mechanistic modeling is based on first principles ([32]), using a priori knowledge of a (often drastically reduced) metabolic network. The resulting reduced reaction scheme can be formulated as follows:

$$\sum_{i \in R_j} k_{i,j} \xi_i \overset{\varphi_j(\xi, \theta_{k,j})}{\rightarrow} \sum_{l \in P_j} k_{l,j} \xi_l \tag{1}$$

where R_j and P_j are respectively the sets of reactants and products of the jth reaction with $j = 1, ..., m$, respectively denoted ξ_i ($i = 1, ..., n_r$) and ξ_l ($l = 1, ..., n_p$). φ_j is the rate of the jth reaction, function of $\xi \in \mathbf{R}^n$, the vector of metabolite concentrations. $\theta_{k,j}$, where the index k stands for "kinetic", is the kinetic parameter vector of the jth reaction, function of the selected kinetic structure. $k_{i,j}$ ($k_{l,j}$) represents the stoichiometric coefficient of the ith (l^{th}) reactant (product) in reaction j. Applying mass balances to (1) leads to the following ordinary differential equation system representing each reaction/product concentration variation with time:

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

where K is the matrix containing the stoichiometric coefficients from (1) and ν is the transport vector, function of the process input/output flows. For the sake of clarity, stoichiometric and kinetic parameters will respectively be denoted by θ_s and θ_k in the following. It should be noticed that unlike chemical reaction schemes which present a unique set of stoichiometric coefficients, biochemical reaction schemes, involving biomass species, may induce some stoichiometric uncertainty where the species consumption and production yields are likely to vary.

MLPCA considers the differential transport-free state vector which, between two consecutive measurements in t_i and t_{i+1}, reads:

$$\xi_{f_i}^{\Delta} = K \int_{t_i}^{t_{i+1}} \varphi(\tau) d\tau \tag{3}$$

where the left-hand side of (3) is composed of the metabolite concentration variations and transport terms, while the right-hand side comprises the stoichiometric basis K and the reaction rates. [19] provides a geometrical interpretation of MLPCA which achieves an approximation of (3) by linear subspaces of increasing dimensions m and returns a $n \times m$ affine subspace basis $\hat{\rho}$ which, by linear combination, is associated to the stoichiometric matrix approximation $\hat{K} = \hat{\rho}G$ with G being a regular $m \times m$ matrix to be defined to impose a maximum of p constraints per column of $\hat{K}$ (for instance, to force $\hat{K}$ to present some normalization with respect to a specific component in the suggested reaction). This association also induces the non-uniqueness of the stoichiometric matrix approximation $\hat{K}$ of rank m. The smallest dimension m providing a sufficiently accurate interpretation of the data can therefore be selected, minimizing a maximum likelihood (ML) cost as follows:

$$J_m = \sum_{i=1}^{n_S} \left(\xi^{\Delta}_{f,meas_i} - \xi^{\Delta,m}_f\right)^T Q_i^{-1} \left(\xi^{\Delta}_{f,meas_i} - \xi^{\Delta,m}_f\right) \tag{4}$$

where n_S is the number of measured samples, $\xi^{\Delta}_{f,meas_i}$ the measured transport-free state vector, Q_i the error covariance matrix accounting for the noise and $\xi^{\Delta,m}_f$, the Maximum Likelihood estimate vector of the m-dimensional model. We select m as the minimal value such that J_m is smaller or equal to the range of a $\chi^2_{n_S \times n}$-distributed random variable (see [19] and [17] for further details).

2.2. STEP 2: identification of the kinetics by lumped flux reconstruction

We assume that the reaction rate structure corresponds to the product of several factors describing occurring activation/saturation and/or inhibition phenomena ([23]). The jth component of φ ($j = 1, ..., m$) is therefore expressed as follows:

$$\varphi^j(t) = \varphi^j_{max} X(t) \prod_{i=1}^{n} h^j_{M,i}\left(\xi_i(t)\right) h^j_{J,i}\left(\xi_i(t)\right) \tag{5}$$

where the index j stands for the reaction number and $\xi_i(t)$ is the concentration of the ith metabolite. φ_{max} is the maximum rate constant. The biomass is assumed to be the auto-catalyst of the rates, driving the overall reaction scheme dynamics. The corresponding concentration X is therefore a factor of each reaction rate. Among the possibly existing kinetic structures, Monod ([33]) and Jerusalimski ([34]) factors are respectively chosen to describe the rate activation and/or inhibition by the ith metabolite, and are denoted:

$$h_{M,i}(t) = \frac{\xi_i(t)}{\xi_i(t) + K_{\xi_i,s}} \tag{6}$$

$$h_{J,i}(t) = \frac{1}{1 + K_{\xi_i,inh}\xi_i(t)} \tag{7}$$

M and J indicating either Monod or Jerusalimski factors (or, more generally, their combined effects in (5)). Metabolites with no kinetic contribution are assimilated to a unitary factor. $K_{\xi_i,s}$ and $K_{\xi_i,inh}$ are respectively the half-saturation and inhibition parameters, assumed to be positive. The kinetic parameter vector can therefore be decomposed as $\theta_k = \left[\varphi_{max}^T \; K_s^T K_{inh}^T\right]$ where φ_{max} K_s and K_{inh} are respectively the maximum rate, the half-saturation and the inhibition constant vectors. In order to identify the kinetic coefficients, several methods may be proposed such as multilinear Gaussian process assumption ([23]), suggesting a Gaussian kernel inspired by the proposed kinetic structure from [35]. However, the latter require the knowledge of the reaction rate measurements which are not assumed to be known a priori. The first step will therefore consist in reconstructing the reaction rates, which are assumed to lump the metabolic fluxes according to the proposed stoichiometric sub-space provided by the MLPCA.

The left-hand side of equation (3) is assumed to be measurable, while the right-hand side comprises the stoichiometric basis obtained by MLPCA and the unknown reaction rate vector. A linear programming problem of the following form may therefore be formulated:

$$\begin{aligned} argmin_x J_{lin} &= b - A\,x \\ s.t.\; x &> 0 \end{aligned} \tag{8}$$

where $b = \frac{d\xi}{dt} - \nu$, $A = K$ and $x = \varphi$. It is assumed that the number of metabolites n is greater than the number of rates m generated by MLPCA, making problem (8) over-determined. A good numerical solution may however be obtained, namely considering a positivity constraint on the solution since the factors (6) and (7) are monotonic positive functions. This statement will be illustrated further in the result section. The resulting measured rate estimates are used in a nonlinear programming problem aiming at identifying the kinetic parameters as follows:

$$argmin_{\theta_k} J = \sum_{j=1}^{m} \sum_{i=1}^{n_S} \left(\varphi_i^j - \hat{\varphi}_i^j\right)^T Q_\varphi^{-1} \left(\varphi_i^j - \hat{\varphi}_i^j\right) - \lambda \sum_{r=1}^{p_k} \theta_k(r) \tag{9}$$

where φ_i^j are the estimates of the measured rates, solutions of problem (8), $Q_\varphi \in \mathbf{R}^{m\times m}$ is the covariance matrix of rate estimation errors and $\hat{\varphi}_i^j$ are the kinetic model estimates, conform to (6) and (7). p_k is the dimension of the kinetic parameter vector θ_k. It should be noticed that the reactant/product labeling of a metabolite may change from one rate to the other, depending on the sign of the corresponding stoichiometric coefficient delivered by the MLPCA. The penalization term weighted by λ aims at providing the smallest order of magnitudes of the kinetic parameter values, forcing the parameters which have no influence on the kinetics to converge to zero and, in turn, reducing their contribution to 1 in (6) and (7).

To assess the local identifiability of the kinetic parameters, depending on the data set informative content, the Fisher Information Matrix (FIM) is computed, over one experiment, as follows:

$$FIM_k = \sum_{i=1}^{n_S} \varphi_{\theta_k}^T(i) Q_\varphi^{-1} \varphi_{\theta_k}(i) \tag{10}$$

where $\varphi_{\theta_k}(i) = \frac{\partial \varphi(i)}{\partial \theta_k}$ is the kinetic parameter sensitivity matrix of dimensions $m \times p_k$. An optimistic estimate of the kinetic parameter estimation error covariance matrix can be inferred from the inverse of the FIM:

$$\hat{P}_k = \hat{\sigma}_k^2 FIM_k^{-1} \tag{11}$$

with $\hat{\sigma}_k^2$ being the a posteriori estimate of the rate measurement error variance obtained from the residual cost function at the optimum:

$$\hat{\sigma}_k^2 = \frac{J}{N_{meas} - p_k} \tag{12}$$

where N_{meas} is the total number of measurements $N_{meas} = n\, n_S$.

The parameter rejection means of this identification step are therefore twofold, considering an upstream parameter set reduction based on a minimum order of magnitude below which the parameter is considered as non influential and a downstream set reduction following the comparison of the parameter estimation error variances $\sigma_{P_k}^2$, read on the diagonal of the covariance matrix P_k, with a maximum level above which local identifiability is insufficient to consider the parameters as significantly influencing the rates. Figure 1 shows the schematized kinetic parameter identification algorithm with two threshold coefficients, ϵ and η, parametrizing the two reduction strategies.

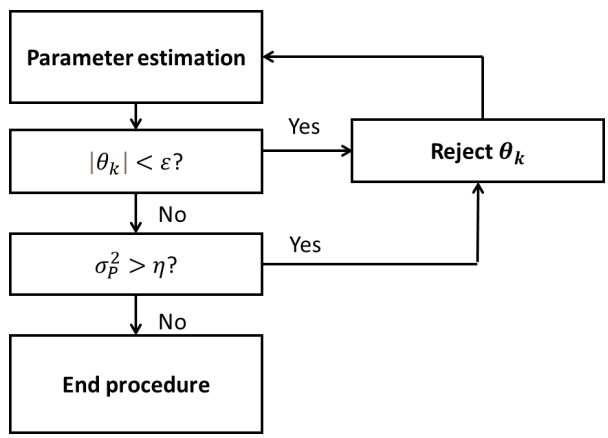

Figure 1: kinetic parameter identification scheme.

2.3. STEP 3: parameter bias cancellation

Identifying the stoichiometric and kinetic parameters separately may unfortunately generate some estimation bias and it is therefore proposed to alleviate this issue by proceeding to a new identification of the full parameter set $\theta_{model} = \begin{bmatrix} \theta & \theta_{IC} \end{bmatrix} = \begin{bmatrix} \theta_s & \theta_k & \theta_{IC} \end{bmatrix}$ where θ_{IC} are the metabolite ordinary differential equation initial conditions, assumed to be uncertain due to the presence of measurement noise. To this end, the following nonlinear ML criterion is minimized:

$$J_{ML} = \sum_{i=1}^{n_S} \left(\xi_{m_i} - \xi_i\right)^T Q_i^{-1} \left(\xi_{m_i} - \xi_i\right) \tag{13}$$

where ξ and ξ_m are respectively the state predictions by the model and the corresponding measurements.

According to step 2, local identifiability is anew assessed in order to detect possible over-parametrization.

The FIM of the full parameter set reads:

$$FIM = \sum_{i=1}^{n_S} \xi_\theta^T(i) Q_i^{-1} \xi_\theta(i) \tag{14}$$

where $\xi_\theta = \frac{\partial \xi}{\partial \theta}$ is the parametric sensitivity of the full parameter set.

The corresponding covariance matrix estimate of the estimation errors becomes:

$$\hat{P} = \hat{\sigma}^2 FIM^{-1} \tag{15}$$

with σ^2 being the a posteriori estimate of the measurement error variance obtained from the residual of J_{ML}:

$$\hat{\sigma}^2 = \frac{J_{ML}}{N_{meas} - n_\theta} \tag{16}$$

where n_θ is the number of estimated parameters.

Even if the above method allows to avoid over-parametrization, parameter non-identifiability may sometimes lead to an ill-conditioned (i.e., not invertible) FIM and, in turn, a non-existing P matrix. In this particular situation, a QR decomposition can be applied in order to rank the parameters in increasing order of dependency ([36]). The most linearly dependent parameter is then detected and rejected before step 3 restarts. The reader may refer to [36] for further information on the use of QR decomposition.

2.4. Data-driven parameter identification algorithm

The proposed global identification scheme imbricating steps 1 to 3 is shown in figure 2 and can be summarized as follows:

- STEP 1: generate the minimal lumped reaction scheme of the provided data set and the corresponding stoichiometric basis;
- STEP 2: based on the calculation of the state fluxes from the data set, calculate the estimates of the rates and identify the kinetic parameters;
- STEP 3: cancel the possible bias by reidentifying the whole parameter set with the parameters from STEP 1 and 2 as initial guesses. Redo STEP 3 as long as the reduction test provides a negative answer or if the user decides to reduce the accuracy tolerance η.

3. Application to animal cell culture process

3.1. Database description

5 experimental data sets are available, describing fed-batch animal cell cultures operated with multi-input feeding of glucose, glutamine and amino acid media. 3 experiments were run with low-concentration amino acid medium and the remaining 2 with high-concentration amino acid medium. Initial concentrations of the metabolites may vary from one experiment to the other, mainly depending on the pre-culture result. For the sake of confidentiality, the metabolite inlet concentrations are not divulged, as well as other information about the operating

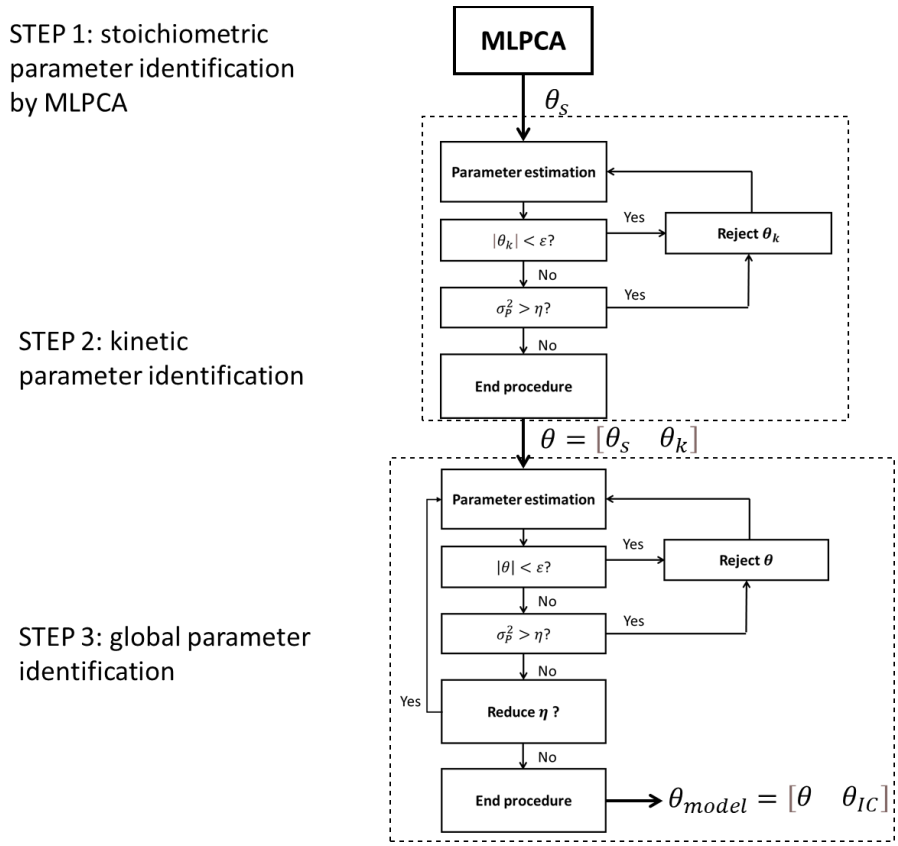

Figure 2: Data-driven model parameter identification scheme.

conditions. The data sets contain the off-line measurements of 25 state variables which are, using a conventional specific numbering:

- The classical metabolites: 1) biomass X, 2) glucose G, 3) lactate L, 4) glutamine Gn and 5) ammonium N;
- The amino acids: 6) aspartic acid Asp, 7) glutaminic acid Glu, 8) serine Ser, 9) asparagine Asn, 10) glycine Gly, 11) histidine His, 12) threonine Thr, 13) arginine Arg, 14) alanine Ala, 15) proline Pro, 16) tyrosine Tyr, 17) cysteine Cys, 18) valine Val, 19) methionine Met, 20) isoleucine Ile, 21) leucine Leu, 22) lysine Lys, 23) phenylalanine Phe, and 24) tryptophan Trp;
- The product of interest: 25) monoclonal antibodies MAb.

The data sets are partitioned in 1 low-concentration and 1 high-concentration amino acid medium data sets dedicated to parameter identification and model direct validation, and 2 low-concentration and 1 high-concentration amino acid medium data sets dedicated to cross-validation. The corresponding data are shown in figure 3.

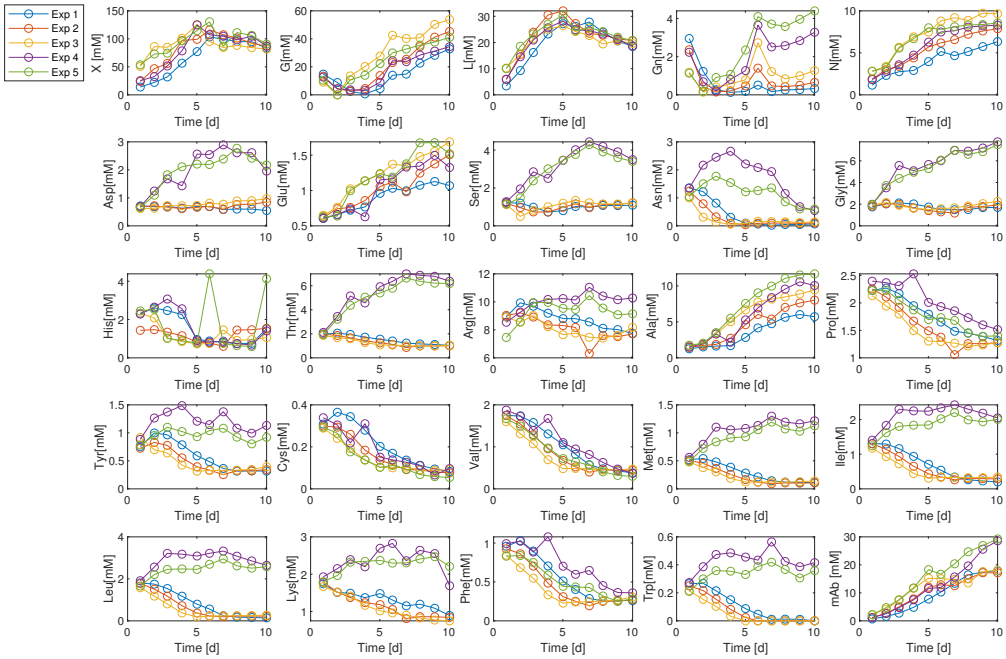

Figure 3: Animal cell fed-batch culture experimental data sets. Experiments 1, 2 and 3 were achieved with low amino acid concentration medium. Experiments 4 and 5 were achieved with high amino acid concentration medium.

3.2. Model validation

3.2.1. STEP 1: MLPCA

Figure 4 shows the log-likelihood function following the application of the methodology defined in STEP 1 to the data sets related to experiments 1 and 4. A 3-dimensional subspace appears to be sufficient to describe the selected data with a probability of $99.5\ \%$ following the χ^2 test.

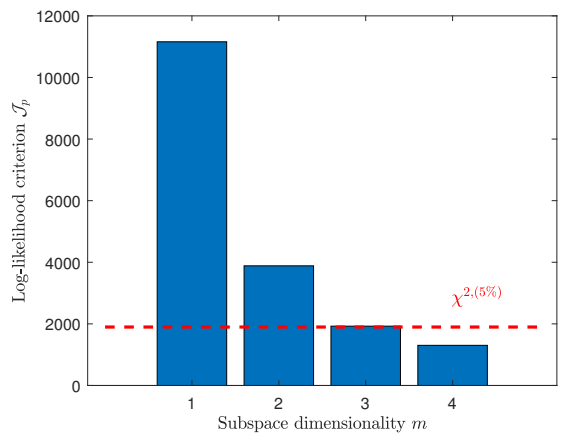

Figure 4: Log-likelihood costs of the p-dimensional subspaces considering a relative noise variance of 10 %. The dashed line represents the χ^2 quantile at 5 %.

The corresponding basis ρ delivered by the MLPCA is used to build a biologically consistent stoichiometric matrix following a linear transformation of the form $\hat{K} = \rho G$ where G allows to impose a maximum of 3 (rank of ρ) specific constraints per column of $\hat{K}$. The following constraints are formulated in a trial to reveal the main metabolic states/pathways which can be observed in the data and which are the overflow metabolism (including glycolysis), biomass death and amino acid metabolism which not explained by the first two reactions :

- The first column (i.e., reaction) of $\hat{K}$ should be normalized with respect to the biomass concentration;
- The second column should be normalized with respect to the biomass concentration and biomass, glucose and glutamine should all present negative or zero stoichiometric coefficients $k_{X,2}, k_{G,2}, k_{Gn,2} \leq 0$;
- The third column should be normalized with respect to the biomass concentration and the ammonium coefficient is zero $k_{N,3} = 0$.

It should be noticed that the normalization with respect to the biomass concentration allows to constrain $\hat{K}$ to be read as the result of a mass balance. The solution to the constrained linear problem provides a G matrix inducing the following stoichiometric matrix:

$$\hat{K} = \rho G = \begin{pmatrix} 1 & -1 & 1 \\ -0.493 & -0.151 & -2.017 \\ 0.266 & -0.443 & -0.153 \\ -0.139 & 0 & -0.466 \\ 0.074 & -0.106 & 0 \\ 0.008 & -0.041 & -0.069 \\ 0.0085 & -0.013 & -0.001 \\ 0.014 & -0.084 & -0.151 \\ 0.003 & -0.099 & -0.222 \\ 0.023 & -0.126 & -0.221 \\ -0.010 & 0.014 & -0.003 \\ 0.012 & -0.105 & -0.207 \\ 0.005 & -0.013 & -0.012 \\ 0.064 & -0.105 & -0.033 \\ -0.006 & 0.004 & -0.011 \\ -0.002 & -0.015 & -0.042 \\ -0.002 & 0.003 & -0.001 \\ -0.011 & 0.011 & -0.009 \\ 0.001 & -0.021 & -0.047 \\ -0.001 & -0.038 & -0.093 \\ -0.002 & -0.059 & -0.147 \\ -0.001 & -0.021 & -0.054 \\ -0.005 & 0.001 & -0.015 \\ -0.001 & -0.010 & -0.025 \\ 0.156 & -0.191 & 0.073 \end{pmatrix} \tag{17}$$

where the state numbering, from row 1 (X) to 25 (MAb) is conform to the data description from section 3.1.

3.2.2. STEP 2: Kinetic parameter identification

The threshold coefficient setting the absolute value below which a kinetic parameter is rejected is 10^{-3}, while the threshold coefficient setting the relative value of the confidence interval above which a kinetic parameter is rejected is 10^5. These intervals correspond to a confidence degree of $95\ \%$ and are therefore calculated, for the ith parameter θ_k as follows:

$$CI = \pm 2\sigma_P \times \frac{100}{\theta_{k,i}} \tag{18}$$

The resulting expressions of the rates are:

$$\varphi^{(1)} = \varphi^{(1)}_{max} X \frac{Cys}{Cys + K_{Cys,s}} \frac{Val}{Val + K_{Val,s}} \frac{Ile}{Ile + K_{Ile,s}} \frac{Leu}{Leu + K_{Leu,s}} \frac{Lys}{Lys + K_{Lys,s}} \frac{K_{Ala,inh}}{Ala + K_{Ala,inh}} \frac{K_{Met,inh}}{Met + K_{Met,inh}} \tag{19a}$$

$$\varphi^{(2)} = \varphi^{(2)}_{max} X \frac{N}{N + K_{N,s}} \frac{Asn}{Asn + K_{Asn,s}} \frac{Arg}{Arg + K_{Arg,s}} \frac{K_{Pro,inh}}{Pro + K_{Pro,inh}} \frac{K_{Val,inh}}{Val + K_{Val,inh}} \tag{19b}$$

$$\varphi^{(3)} = \varphi^{(3)}_{max} X \tag{19c}$$

From these latter reaction rate expressions, it is shown that some essential amino acids (except Cys which can however be assimilated as a conditionally non-essential amino acid) are detected as responsible of the activation of the first reaction, assumed to represent the main substrate consumption pathway (including glycolysis), which makes sense. This first reaction is also a priori inhibited by Met and Ala, respectively essential and non-essential. The second reaction, modeling the cell death rate, is logically activated by the presence of N, which is recognized to be a cell growth inhibitor at significant concentration levels. Asn and Arg, which are non-essential, also activate cell death when accumulated while Pro and Val conversely inhibit it. It should be noticed that no activation/inhibition mechanism is detected regarding the third rate.

Parameter values are shown in Table 1.

3.2.3. STEP 3: Bias cancellation

This last step considers the estimations from the previous steps as initial guess as well as the identification of the ordinary differential equation initial conditions. For the sake of illustration, two runs of the procedure are achieved with confidence interval thresholds (η) of 10^4 and 10^3 and $\epsilon = 10^{-3}$ while observing the effect on $\hat{K}$, rates (19) and the cost function residual J_{ML}. Tables 2 to 5 show the parameter values and the corresponding relative confidence intervals at $95\ \%$ for each step. The cost function residuals for $\eta = 10^4$ and 10^3 are respectively 0.941 and 0.834.

Parameter values and their relative confidence intervals are shown in Tables 2 to 5. In overall, the values from Tables 2 and 3 are very close and the main difference lies in the less amount of parameters (especially kinetic parameters)

Table 1: Kinetic parameter values following STEP 2

Parameter	Value	Unit
$\varphi_{max}^{(1)}$	17.166	d^{-1}
$\varphi_{max}^{(2)}$	199.847	d^{-1}
$\varphi_{max}^{(3)}$	0.0243	d^{-1}
$K_{Cys,s}$	0.099	mM
$K_{Val,s}$	0.288	mM
$K_{Ile,s}$	0.294	mM
$K_{Leu,s}$	0.298	mM
$K_{Lys,s}$	2.805	mM
$K_{N,s}$	48.147	mM
$K_{Asn,s}$	0.135	mM
$K_{Arg,s}$	67.467	mM
$K_{Ala,inh}$	0.113	mM^{-1}
$K_{Met,inh}$	6.179	mM^{-1}
$K_{Pro,inh}$	19.991	mM^{-1}
$K_{Val,inh}$	0.221	mM^{-1}

considered by the second model (with $\eta = 10^3$), to the cost of higher kinetic uncertainties as highlighted by Table 5. Moreover, a majority of parameters from reaction 3, in both cases, are poorly identifiable, which, at first sight, would suggest to reduce the model to 2 reactions, in opposition to the results generated by the MLPCA. Choosing $\eta = 10^2$ rejects the last statement (for the sake of conciseness and clarity, the corresponding results are not provided but the direct validation is shown in Figures 5 and 6). Indeed, in the latter case, parameters from reaction 3 are all rejected but the cost function dramatically increases ($J_{ML} = 8.694$), leading to a bad fitting. Three reactions are therefore necessary even if the third one is poorly identifiable.

Table 2: Final model stoichiometric parameter values.

Parameter	$\eta = 10^4$	$\eta = 10^3$	Unit
$k_{G,1}$	-0.264	-0.239	$mmol/10^9 cells$
$k_{G,3}$	-0.090	-0.145	$mmol/10^9 cells$

$k_{L,1}$	0.374	0.300	$mmol/10^9 cells$
$k_{L,2}$	-0.220	-0.247	$mmol/10^9 cells$
$k_{Gn,1}$	-0.071	-0.061	$mmol/10^9 cells$
$k_{Gn,3}$	-0.014	-0.031	$mmol/10^9 cells$
$k_{N,1}$	0.045	0.043	$mmol/10^9 cells$
$k_{N,2}$	-0.002	-0.008	$mmol/10^9 cells$
$k_{Asp,1}$	0.003	0.003	$mmol/10^9 cells$
$k_{Asp,2}$	-0.003	-0.004	$mmol/10^9 cells$
$k_{Glu,1}$	0.005	0.004	$mmol/10^9 cells$
$k_{Ser,2}$	-0.006	-0.007	$mmol/10^9 cells$
$k_{Asn,3}$	-0.057	-0.113	$mmol/10^9 cells$
$k_{Gly,3}$	-0.001	-0.002	$mmol/10^9 cells$
$k_{His,1}$	-0.019	-0.018	$mmol/10^9 cells$
$k_{His,2}$	0.008	0.011	$mmol/10^9 cells$
$k_{Thr,3}$	-0.009	-0.018	$mmol/10^9 cells$
$k_{Arg,1}$	0.005	0.004	$mmol/10^9 cells$
$k_{Arg,3}$	-0.008	-0.017	$mmol/10^9 cells$
$k_{Ala,1}$	0.042	0.038	$mmol/10^9 cells$
$k_{Pro,1}$	-0.007	-0.006	$mmol/10^9 cells$
$k_{Tyr,1}$	-0.006	-0.005	$mmol/10^9 cells$
$k_{Cys,1}$	-0.001	-0.001	$mmol/10^9 cells$
$k_{Cys,2}$	0.001	0.001	$mmol/10^9 cells$
$k_{Cys,3}$	-0.003	-0.004	$mmol/10^9 cells$
$k_{Val,1}$	-0.008	-0.008	$mmol/10^9 cells$
$k_{Val,2}$	0.003	0.003	$mmol/10^9 cells$
$k_{Val,3}$	-0.006	-0.009	$mmol/10^9 cells$
$k_{Met,3}$	-0.005	-0.011	$mmol/10^9 cells$
$k_{Ile,1}$	-0.008	-0.008	$mmol/10^9 cells$
$k_{Leu,1}$	-0.014	-0.015	$mmol/10^9 cells$
$k_{Lys,2}$	-0.005	-0.006	$mmol/10^9 cells$
$k_{Phe,1}$	-0.006	-0.005	$mmol/10^9 cells$
$k_{Trp,1}$	-0.003	-0.003	$mmol/10^9 cells$
$k_{MAb,1}$	0.051	0.046	$mmol/10^9 cells$
$k_{MAb,3}$	0.115	0.219	$mmol/10^9 cells$

Table 3: Final model kinetic parameter values.

Parameter	$\eta = 10^4$	$\eta = 10^3$	Unit
$\varphi_{max}^{(1)}$	3.160	2.248	d^{-1}
$\varphi_{max}^{(2)}$	2.087	3.121	d^{-1}
$\varphi_{max}^{(3)}$	0.233	0.117	d^{-1}
$K_{Cys,s}$	0.083	0.489	mM
$K_{Ile,s}$	0.750	0.276	mM
$K_{Leu,s}$	0.032	-	mM
$K_{N,s}$	23.786	55.755	mM
$K_{Ala,inh}$	0.536	0.032	mM^{-1}
$K_{Val,inh}$	0.168	-	mM^{-1}

Table 4: Final model stoichiometric parameter relative confidence intervals in $\%$.

Parameter	$\eta = 10^4$	$\eta = 10^3$	Unit
$k_{G,1}$	37.517	27.374	$\%$
$k_{G,3}$	166.863	188.586	$\%$
$k_{L,1}$	28.229	20.896	$\%$
$k_{L,2}$	49.048	35.254	$\%$
$k_{Gn,1}$	35.987	24.304	$\%$
$k_{Gn,3}$	223.707	191.442	$\%$
$k_{N,1}$	34.446	21.996	$\%$
$k_{N,2}$	460.776	148.788	$\%$
$k_{Asp,1}$	111.359	96.359	$\%$
$k_{Asp,2}$	108.748	92.770	$\%$
$k_{Glu,1}$	31.998	19.973	$\%$
$k_{Ser,2}$	70.721	58.520	$\%$
$k_{Asn,3}$	105.565	115.967	$\%$
$k_{Gly,3}$	409.795	453.266	$\%$
$k_{His,1}$	31.066	22.159	$\%$
$k_{His,2}$	62.114	45.153	$\%$
$k_{Thr,3}$	118.061	126.863	$\%$

$k_{Arg,1}$	335.723	337.927	%
$k_{Arg,3}$	305.202	302.173	%
$k_{Ala,1}$	30.996	18.669	%
$k_{Pro,1}$	33.516	22.065	%
$k_{Tyr,1}$	34.311	23.090	%
$k_{Cys,1}$	49.178	37.697	%
$k_{Cys,2}$	65.383	55.129	%
$k_{Cys,3}$	91.336	118.755	%
$k_{Val,1}$	47.773	38.926	%
$k_{Val,2}$	186.624	164.861	%
$k_{Val,3}$	195.465	211.564	%
$k_{Met,3}$	107.418	118.034	%
$k_{Ile,1}$	36.778	25.868	%
$k_{Leu,1}$	34.500	23.283	%
$k_{Lys,2}$	59.242	46.947	%
$k_{Phe,1}$	31.796	19.708	%
$k_{Trp,1}$	35.200	23.866	%
$k_{MAb,1}$	68.812	63.661	%
$k_{MAb,3}$	114.903	127.403	%

Table 5: Final model kinetic parameter relative confidence intervals in %.

Parameter	$\eta = 10^4$	$\eta = 10^3$	Unit
$\varphi_{max}^{(1)}$	45.131	163.864	%
$\varphi_{max}^{(2)}$	216.767	489.383	%
$\varphi_{max}^{(3)}$	104.189	116.388	%
$K_{Cys,s}$	324.653	576.445	%
$K_{Ile,s}$	49.039	389.985	%
$K_{Leu,s}$	112.007	-	%
$K_{N,s}$	293.787	319.743	%
$K_{Ala,inh}$	116.253	104.292	%
$K_{Val,inh}$	770.766	-	%

Figures 5 and 6 respectively show the direct validations of the models generated with the different η values, on experiments 1 (with the first medium) and 5 (with the second medium). Obviously, the fitting is almost the same for $\eta = 10^4$ and 10^3.

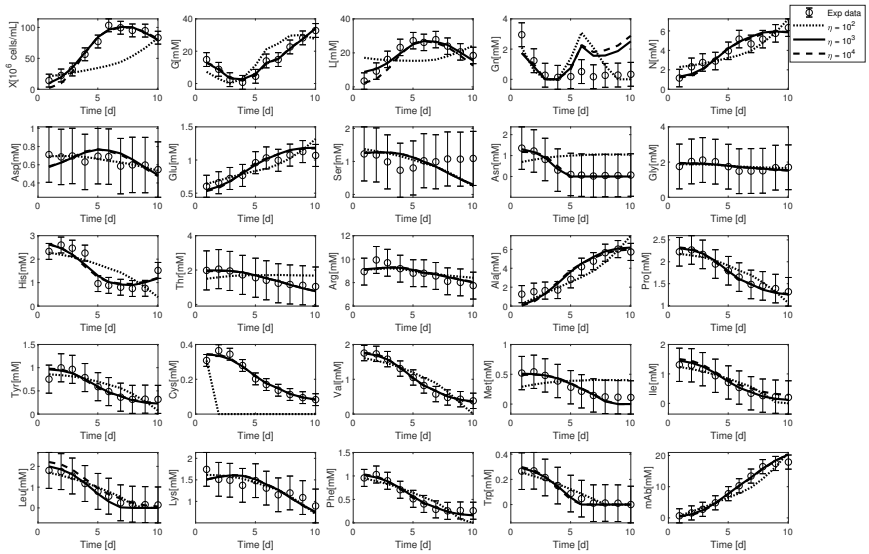

Figure 5: Direct validation of the proposed models for different η values on experiment 1 (first medium). Dotted, continuous and dashed lines respectively correspond to $\eta = 10^2$, $\eta = 10^3$ and $\eta = 10^4$ while the experimental data are represented by the errorbars.

Choosing one of the two presented models therefore appears to be the best compromise. In order to distinguish the best candidate, based on their predictive capabilities, we proceed to the cross-validations on the 3 remaining data sets.

3.2.4. Model cross-validation

During the cross-validations, only initial conditions are identified while model parameters are set to the results of the direct validations. The 3 remaining fed-batch experiments (2 with low and 1 with high amino acid concentration media) are used in this purpose and the fitting results are shown in Figures 7 to 9 for the two selected data-driven models.

In order to ease the comparison, Table 6 summarizes the model fitting cost residuals for each cross-validation experiment.

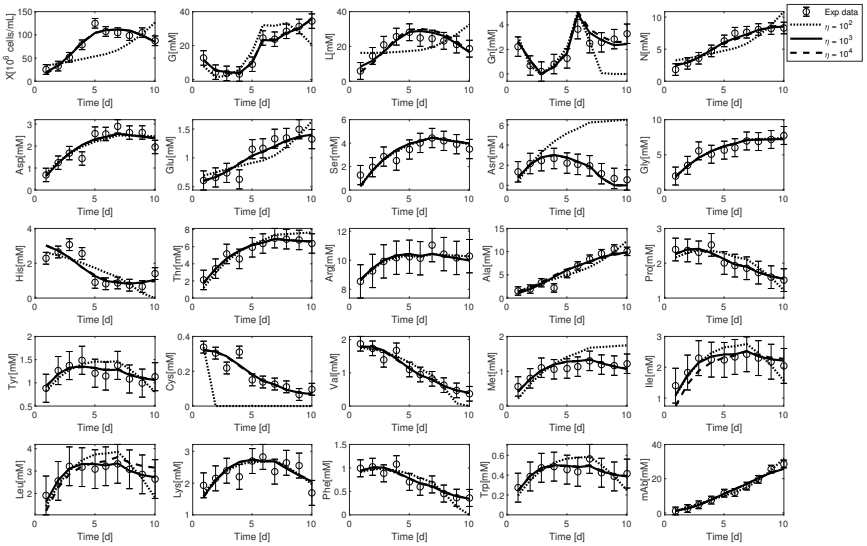

Figure 6: Direct validation of the proposed models for different η values on experiment 2 (second medium). Dotted, continuous and dashed lines respectively correspond to $\eta = 10^2$, $\eta = 10^3$ and $\eta = 10^4$ while the experimental data are represented by the errorbars.

Table 6: Data-driven model fitting cost function J_{ML} residuals during cross-validations.

Experiment	J_{ML} ($\eta = 10^4$)	J_{ML} ($\eta = 10^3$)
3	0.553	0.444
4	1.076	0.879
5	0.974	0.840

From a qualitative point of view, the predictive capability of the data-driven model with $\eta = 10^3$ is better than the second model, which is confirmed by the residuals from Table 6. Experiments 3 and 4 show some bad fittings regarding glutamine and serine concentrations, which is the result of an amplified phenomenon already observed in experiment 1 (direct validation). However, experiment 5 proposes the best results, namely for $\eta = 10^3$. The difference of fitting quality between the selected experiments seems to be connected to the used medium. A first element of explanation could be the higher concentration of glutamine in the second medium, combined to the impossibility to switch from one rate to another but

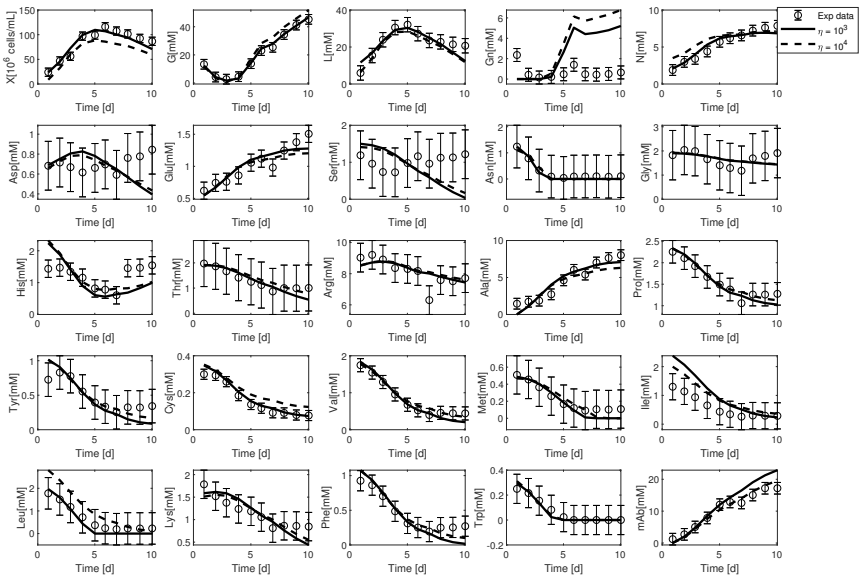

Figure 7: Cross-validation of the proposed models for different η values on experiment 3 (first medium). Continuous and dashed lines respectively correspond to $\eta = 10^3$ and $\eta = 10^4$ while the experimental data are represented by the errorbars.

to continuously consider all rates, leading to an undesired accumulation of glutamine in the model prediction. The final reaction scheme is written as follows:

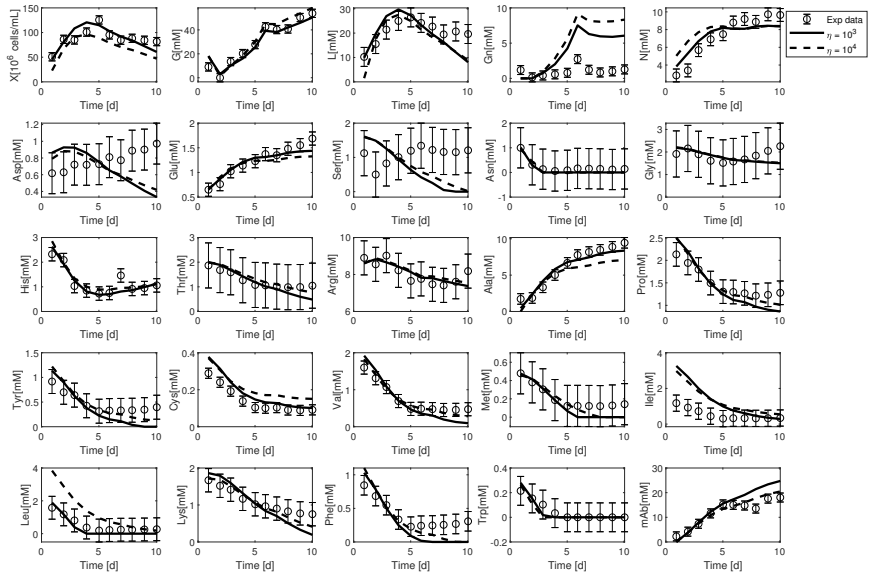

Figure 8: Cross-validation of the proposed models for different η values on experiment 4 (first medium). Continuous and dashed lines respectively correspond to $\eta = 10^3$ and $\eta = 10^4$ while the experimental data are represented by the errorbars.

Reaction 1 :

$$k_{G,1}\mathrm{G} + k_{Gn,1}\mathrm{Gn} + \sum_{AA_{r,1}} k_{AA_{r,1}}\mathrm{AA_{r,1}} \tag{20a}$$
$$\overset{\varphi^{(1)}}{\rightarrow} \mathrm{X} + k_{L,1}\mathrm{L} + k_{N,1}\mathrm{N} + k_{MAb,1}\mathrm{MAb} + \sum_{AA_{p,1}} k_{AA_{p,1}}\mathrm{AA_{p,1}}$$

Reaction 2 :

$$\mathrm{X} + k_{L,2}\mathrm{L} + k_{N,2}\mathrm{N} + \sum_{AA_{r,2}} k_{AA_{r,2}}\mathrm{AA_{r,2}} \overset{\varphi^{(2)}}{\rightarrow} \sum_{AA_{p,2}} k_{AA_{p,2}}\mathrm{AA_{p,2}} \tag{20b}$$

Reaction 3 :

$$k_{G,3}\mathrm{G} + k_{Gn,3}\mathrm{Gn} + \sum_{AA_{r,3}} k_{AA_{r,3}}\mathrm{AA_{r,3}} \overset{\varphi^{(3)}}{\rightarrow} \mathrm{X} + k_{MAb,3}\mathrm{MAb} \tag{20c}$$

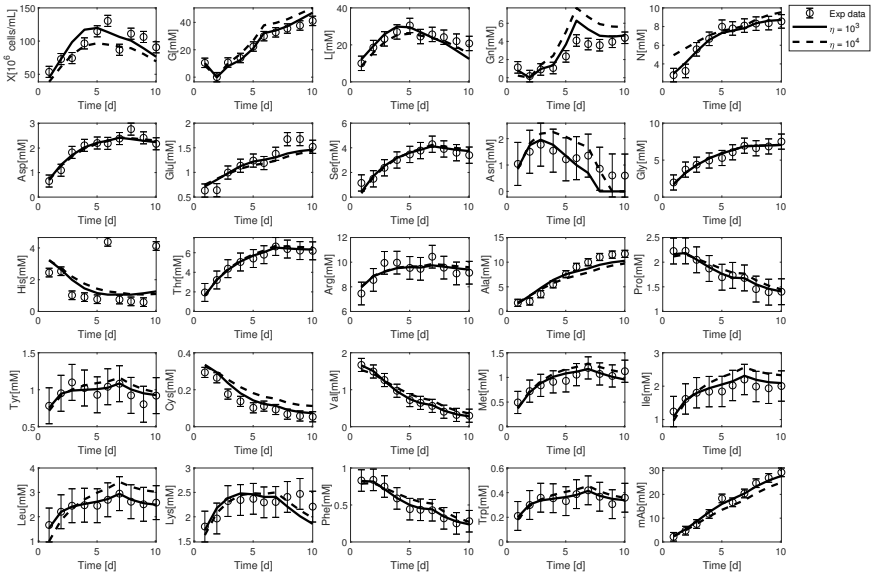

Figure 9: Cross-validation of the proposed models for different η values on experiment 5 (second medium). Continuous and dashed lines respectively correspond to $\eta = 10^3$ and $\eta = 10^4$ while the experimental data are represented by the errorbars.

where indices AA, r and p respectively stand for amino acids, reactants and products. In order to provide a clearer view of the roles of the amino acids in these reactions, Table 7 gathers the several amino acids per group of reactants/products per reaction, while the kinetics read:

$$\varphi^{(1)} = \varphi^{(1)}_{max} X \frac{Cys}{Cys + K_{Cys,s}} \frac{Ile}{Ile + K_{Ile,s}} \frac{K_{Ala,inh}}{Ala + K_{Ala,inh}} \tag{21a}$$

$$\varphi^{(2)} = \varphi^{(2)}_{max} X \frac{N}{N + K_{N,s}} \tag{21b}$$

$$\varphi^{(3)} = \varphi^{(3)}_{max} X \tag{21c}$$

The first reaction can be assimilated to the overflow metabolism since the main substrate metabolites are consumed, leading to the production of lactate, ammonium and other non-essential amino acids such as aspartic acid, glutaminic

Table 7: Assumed amino acid roles in the reaction scheme (20)

Reaction	Reactants	Products
1	His, Pro, Tyr, Cys, Val, Ile, Leu, Phe, Trp	Asp, Glu, Arg, Ala
2	Asp, Ser, Lys	His, Cys, Val
3	Asn, Gly, Thr, Arg, Cys, Val, Met	/

acid, arginine and alanine. This reaction is activated by the presence of cysteine and isoleucine (essential amino acids) while inhibited by the presence of alanine (non-essential). The second reaction lumps biomass decay as well as lactate reconsumption metabolism, and is activated by the presence of ammonium. Aspartic acid (non-essential), serine (conditionally non-essential) and lysine (essential) are consumed and histidine (essential), cysteine (non-essential) and valine (essential) are produced. Eventually, the third reaction contributes to the consumption of metabolites such as glucose, glutamine and other essential or conditionally non-essential amino acids, except cysteine, which is non-essential. This could be interpreted as the lumping of glycolysis and part of the amino acid metabolism, which seems to be correlated.

Of course, the latter interpretations are subject to discussion since they rise from a data-driven macroscopic model therefore considering only the information from the datasets and a limited number of reactions lumping all the cell metabolic fluxes. However, the usefulness of the presented results so far is illustrated in the next section, proposing the robustness assessment of control performances based on the selected data-driven model when facing possible structural mismatch with the true plant, possibly explaining the current glutamine model/data mismatch.

4. Multi-stage nonlinear model predictive control

Following the biologically inherent uncertain metabolic behavior of the cells, a robust control framework must be considered, not only taking parameter variations into account but also possible model/plant structure mismatch. Multi-stage nonlinear model predictive control (MS-NMPC) has been proposed by [37] as a non-conservative scheme considering multiple scenarios related to parameter uncertainty, provided that the latter is well estimated. [13] have recently shown that the method is also able to deal with structural mismatch issues, proposing scenarios switching between several reduced metabolic candidate models. We propose an extension of this principle, not exactly to a multiple model case, but to possi-

ble data-driven modeled rate activation/deactivation in conformity with the bottleneck assumption of [5]. Model (20) indeed lumps numerous metabolic pathways in three corresponding macroscopic reactions labeled as substrate oxidation and overflow metabolism (reaction 1), cell starving inducing lactate reconsumption obviously correlated to cell decay (reaction 2) and a remaining third reaction in which unconsidered consumptions and productions of metabolites from reactions 1 and 2 are gathered in a coarse amino acid metabolism (reaction 3). However, the proposed methodology is not able to state on the possible discontinuous activation/deactivation since all rates are continuously formulated. A scenario tree based on the separate existences of reactions 1 and 2 is proposed. Figure 10 illustrates this proposition where a robust horizon H_r of 2 sampling times is considered, either choosing to follow the overflow (reactions 1 and 3) or starving (reactions 2 and 3) macroscopic pathways.

$$J = \sum_{i=1}^{2^r} J_i$$

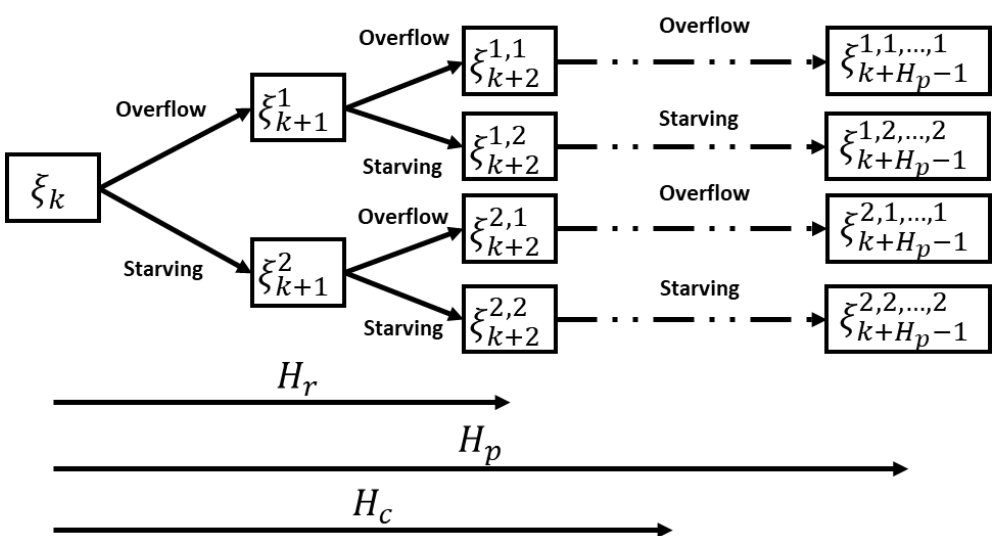

Figure 10: Scenario tree of the proposed MS-NMPC strategy. Two scenarios are considered at each sampling time, based on model 20: overflow (reactions 1 and 3) or (reactions 2 and 3). H_p, H_c and H_r respectively stand for the prediction, control and robust horizons.

We consider an optimal control problem where the monoclonal antibody (MAb)

productivity is maximized (i.e., the production over a specific fixed time) while regulating the glucose (G) and glutamine (Gn) concentration levels. To this end, three input actuators are considered under the form of peristaltic pumps delivering glucose, glutamine and the amino acid medium, separately. The following objective function is therefore considered for all scenarios:

$$\Phi = \sum_{i=1}^{H_p} -MAb(t_i) + \alpha \left(G(t_i) - G_{ref}(t_i)\right)^2 + \beta \left(Gn(t_i) - Gn_{ref}(t_i)\right)^2 \quad (22)$$

where α and β are weighting parameters allowing to adapt the tracking behavior of the controller. G_{ref} and Gn_{ref} are respectively the glucose and glutamine reference concentrations. Expression (22) represents the distances of G and Gn from their respective reference levels, which should be minimized over a specific horizon H_p while the MAb concentration should be maximized (i.e., $-MAb$ should be minimized).

Eventually, the optimal control problem can be formulated as follows:

$$min_{\underline{u_i^j}} \sum_j \omega_j \Phi_j \left(\xi^j, \underline{u^j}\right) \quad (23a)$$

$$s.t. \quad \dot{\xi^j} = \hat{K}\varphi(\xi^j) + \nu(\xi^j, \underline{u^j}) \quad (23b)$$

$$u(t) = u(t_{i+H_c-1}), \quad t \in \left[t_{i+H_c}, t_{i+H_p-1}\right] \quad (23c)$$

$$u_L \leq u \leq u_U \quad (23d)$$

$$u_L \leq u_c \leq u_U, c = 1, ..., H_c \quad (23e)$$

$$\xi_L^p \leq \xi_L^p \leq \xi_U^p, p = 1, ..., H_p \quad (23f)$$

where H_c and H_p are, respectively, the control and prediction horizons and (23b) is the model obtained from the application of mass balance to (20) for a specific scenario j. The constrained problem (23) consists in minimizing a sum of objective functions Φ_j which are calculated as in (22) for each possible scenario occurring with a probability reflected by ω_j. This constrained nonlinear programming problem is solved in the following by multiple shooting using the 'fmincon' solver from Matlab®. The MS-NMPC calculates an open-loop optimal trajectory, updated by the closed-loop activated at each sampling time, when new measurements are available.

5. Numerical results

A multi-input multi-output (MIMO) case is considered with 3 pumps controlling inlet flows of amino acids F_{aa}, glucose F_G and glutamine F_{Gn}, separately (for confidentiality purpose, the concentrations are not revealed). All state variables are supposed to be measured. This last statement is indeed practically difficult to achieve but additional software tools such as observers could be used to estimate possibly unmeasurable state variables as well as data-driven model reductions (with less state variables). The main purpose of the following numerical simulations is to assess the feasibility of the proposed control policy. The imposed constraints are shown in Table 8 where $u = [F_{aa}, F_G, F_{Gn}]$. These ones represent physical limitations of the pumps supposed to work in the range $[0\ 0.1]\ L\ d^{-1}$. Moreover, a corridor is imposed on glucose and glutamine concentrations to ensure that criterion (22), where α and β are respectively set to 5 and 10 (by trial and error), regulates G and Gn sufficiently close to the imposed references $G_{ref} = 1.5\ g\ L^{-1}$ and $Gn_{ref} = 1\ g\ L^{-1}$. For the sake of realism, a relative white noise level with $3\ \%$ standard deviation is added on the state variable measurements.

Table 8: NMPC constraints

Variable	Lower bound	Upper bound	Unit
F_{aa}	0	0.1	$L\ d^{-1}$
F_G	0	0.1	$L\ d^{-1}$
F_{Gn}	0	0.1	$L\ d^{-1}$
G	1	3	$g\ L^{-1}$
Gn	0.1	2	$g\ L^{-1}$

Coefficients ω_j should be set based on an estimation of their occurrence likelihood. We propose the following weight distribution : $\omega_1 = 2$, $\omega_2 = 0.1$, $\omega_3 = 0.1$ and $\omega_4 = 1$, where, based on the representation from Figure 10, the overflow scenario 1 is the most likely to occur, the starving scenario 4 is less probable and switching scenarios 2-3 should be almost punctual. It should be noticed that the latter representation is inspired from the classical progress of a fed-batch culture.

Applications of multivariable economic NMPC, considering criterion (22) with a sole scenario (i.e., where all rates are likely to occur at the same time) and the multi-stage formulation as previously described, are carried out in simulation. The assumed true plant is simulated by the data-driven model with switching modes.

Only reactions 1 and 3 occur until $t = 4\ d$. Then, a switch to reactions 2 and 3 (starving mode) takes place and the first reaction rate is zero until the final culture time ($t = 10\ d$). The resulting plant/model mismatch is assumed to be structural, justifying the use of the multi-stage formulation. The operating parameters are sequentially selected by trial and error, based on the economic NMPC problem, assessing the best sampling period T_s while increasing prediction and control horizons from 2 to 4 times T_s. The considered qualitative and quantitative criteria to select the best NMPC configuration are (i) the possible violation of the imposed constraints, (ii) the glucose and glutamine regulation root mean-square errors (RMSE) and (iii) the level of monoclonal antibody production. The results are summarized in Tables 9 and 10, highlighting, for all proposed horizon configurations, the sampling time, the constraint violation confirmation, the maximum computing time to solve one NMPC problem iteration (which should remain lower than T_s for the sake of practical feasibility), the glucose and glutamine regulation RMSEs and the final mAb level. These numerical results were generated on Matlab®with an Intel(R) Xeon(R) CPU with two processors $E5620\ 2.4\ GHz$ and $24\ Go$ random-access memory (RAM).

Table 9: Economic NMPC configuration results

$H_c = H_p = 2$						
$T_S[d]$	$-[s]$	Cons. Viol.	Max time (s)	$RMSE_G\ (mM)$	$RMSE_{Gn}\ (mM)$	$mAb_{t_f}\ (mM)$
0.02	1728	Yes	98	0.173	0.117	41.712
0.04	3456	No	94	0.198	0.132	14.268
0.06	5184	No	98	0.320	0.118	10.554
0.08	6912	No	98	0.276	0.094	17.13
0.1	8640	No	103	0.662	0.200	11.271
/	/	/	100.271	0.278	0.142	16.497
$H_c = H_p = 3$						
$T_S[d]$	$-[s]$	Cons. Viol.	Max time (s)	$RMSE_G\ (mM)$	$RMSE_{Gn}\ (mM)$	$mAb_{t_f}\ (mM)$
0.02	1728	No	174	0.130	0.273	11.283
0.04	3456	No	178	0.263	0.080	7.902
0.06	5184	Yes	215	0.365	0.176	8.807
0.08	6912	Yes	195	0.333	0.591	13.126
0.1	8640	No	207	0.660	0.194	15.167
/	/	/	193.8	0.350	0.263	11.257

$H_c = H_p = 4$						
$T_S[d]$	$-[s]$	Cons. Viol.	Max time (s)	$RMSE_G\ (mM)$	$RMSE_{Gn}\ (mM)$	$mAb_{t_f}\ (mM)$
0.02	1728	No	361	0.197	0.376	15.861
0.04	3456	Yes	345	0.231	0.115	19.207
0.06	5184	No	291	0.486	0.1501	29.259
0.08	6912	Yes	344	0.399	0.998	15.338
0.1	8640	No	288	0.455	0.147	12.497
/	/	/	325.8	0.353	0.357	18.432

Table 10: Multi-stage NMPC configuration results

$H_c = H_p = 2$						
$T_S[d]$	$-[s]$	Cons. Viol.	Max time (s)	$RMSE_G\ (mM)$	$RMSE_{Gn}\ (mM)$	$mAb_{t_f}\ (mM)$
0.02	1728	Yes	320	0.479	0.151	13.517
0.04	3456	No	250	0.1983	0.122	18.639
0.06	5184	No	300	0.185	0.115	13.273
0.08	6912	Yes	305	0.455	0.899	21.721
0.1	8640	No	315	0.326	0.139	11.469
/	/	/	313	0.317	0.289	15.226
$H_c = H_p = 3$						
$T_S[d]$	$-[s]$	Cons. Viol.	Max time (s)	$RMSE_G\ (mM)$	$RMSE_{Gn}\ (mM)$	$mAb_{t_f}\ (mM)$
0.02	1728	Yes	772	0.277	0.343	21.435
0.04	3456	No	707	0.284	0.138	16.682
0.06	5184	Yes	734	1.452	0.166	12.102
0.08	6912	Yes	711	1.036	0.324	10.338
0.1	8640	No	3793	0.257	0.147	17.638
/	/	/	1315	0.647	0.190	15.440
$H_c = H_p = 4$						
$T_S[d]$	$-[s]$	Cons. Viol.	Max time (s)	$RMSE_G\ (mM)$	$RMSE_{Gn}\ (mM)$	$mAb_{t_f}\ (mM)$
0.02	1728	No	3800	0.194	0.182	21.718
0.04	3456	No	3993	0.243	0.128	14.491
0.06	5184	Yes	3620	0.399	0.999	9.369
0.08	6912	Yes	2500	1.097	0.871	16.174

0.1	8640	Yes	1230	0.519	0.999	17.099
/	/	/	3028.6	0.490	0.636	15.770

In overall, increasing the sampling time deteriorates the performances of both MPC strategies in correlation with the increasing computing times. The latter remark therefore also applies to the prediction horizon. This observation could be generalized as the more reactive the NMPC can be, i.e. the smaller T_s, the better the tracking results become. However, for very small sampling times and horizons (i.e., $T_s = 0.02\ d$ and $H_p = H_c = 0.04\ d$), the tracking becomes more chaotic as well as the actuator solicitations which may lead to constraint violations. The best results are obtained for $T_s = 0.04\ d$ for both strategies with better performances of the MS-NMPC. The latter indeed proposes smoother input trajectories and highlights better tracking performances (RMSEs), even if providing an acceptable but lower final mAb concentration as a tradeoff. The MSNMPC therefore requires a small sampling time to be sufficiently reactive but also a sufficient horizon ($H_p = 0.08$ for $T_s = 0.04\ d$) to keep some robustness with respect to structural model changes. The results from [27] supports the previous statements, namely regarding the choice of relatively short horizons (ranging from 3 to 6 sampling times) for a bacteria model with only 6 state variables. Since the current mammalian cell data-driven model counts 25 state variables and requires an important number of optimization runs following the multi-stage scenarios, the reduction of the corresponding horizons seems legitimate.

A qualitative estimation of these best results is shown in Figures 11 and 12. Almost the same behaviors are adopted by the classical and multi-stage predictive controllers, except in the first days, where the classical NMPC is suddenly close to the lower bound on glutamine, while the MSNMPC is able to keep the same signal in a safer range. Figure 12 also shows that the MSNMPC performance requires faster but small variations of the actuators while the classical NMPC presents smoother input trajectories with, however, more important peak variations during the transient phase of the first two days.

In summary, the results from Tables 9 and 10 show that the performances of both MPC algorithms depend on the sampling time T_s and the prediction horizon H_p, which should be carefully selected. When observing the controller behaviors in a specific and favorable configuration presenting good RMSEs and MAb productions, the MSNMPC highlights more robustness with respect to the structural

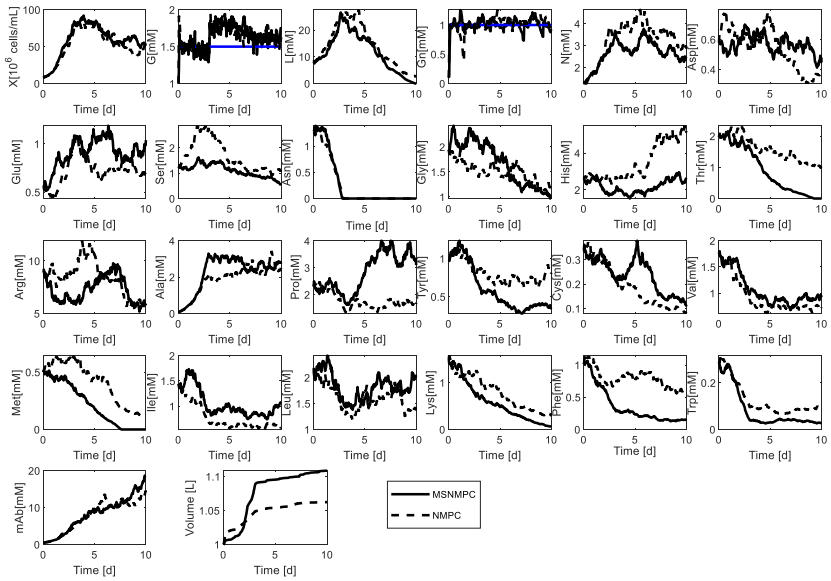

Figure 11: State evolutions of the NMPC simulation. The tracking references are represented by the continuous blue lines.

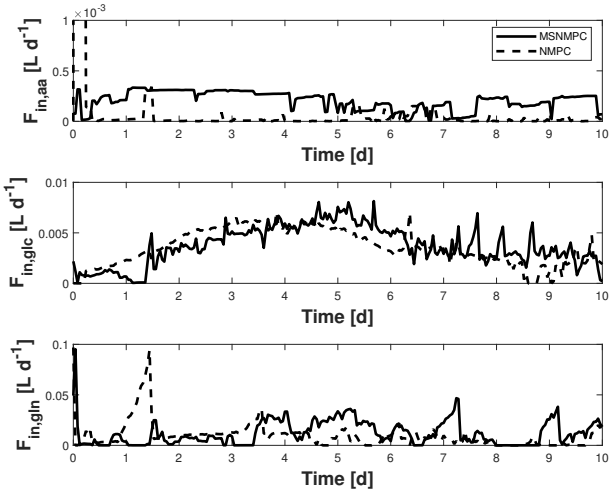

Figure 12: Feed rate evolutions of the NMPC simulations.

uncertainties. Indeed, the regulated state variables as well as input trajectories of the MSNMPC present a smaller number of important peak variations.

The availability of the suggested on-line monitoring of all state variables is not straightforward but some recent research advancements on chemometric models using NIR or fluorescence spectra coupled to software sensors are promising in view of experimental implementations, as suggested in [38], [39], [28] and [40]. Moreover, the data-driven model kinetics suggest that only a few amino acids drive the process in the selected operating conditions. Advanced monitoring devices should therefore first focus on this limited amount of metabolites, increasing the chances of practical feasibility.

6. Conclusions

This paper presents a data-driven modeling procedure applied, in a predictive control framework, to mammalian cell fed-batch culture process. Based on selected datasets, a macroscopic reaction scheme with minimal complexity (i.e., with a minimal number of macroreactions) is determined applying maximum likelihood principal component analysis (MLPCA), also providing the stoichiometry. The latter results are then exploited by a kinetic generation method which allows to identify the metabolites driving the macroreactions. In order to remove the possible estimation bias following the separated parameter identifications, a global nonlinear identification procedure, using as initial guess the parameter values obtained so far, is achieved, also aiming at sequentially removing any parameter which would present poor identifiability, based on the Fisher Information Matrix. The obtained mechanistic model presents continuous reaction rates and is not able to detect possible metabolic switches successively canceling and activating some of the rates. A robust nonlinear model predictive control (NMPC) formulation, based on multi-stage predictions, is therefore proposed. The latter is challenged and compared to the classical economic NMPC, in an application to a simulated plant with discontinuous kinetics modeling the metabolic switches. The results show that the classical NMPC already performs well but the MSNMPC globally exhibits a more accurate tracking for acceptable productivity levels. Future work includes the combination of the proposed data-driven method with chemometric models and software sensors, in order to improve the monitoring aspects. Experimental validations are also considered as short-term important perspectives.

Acknowledgments

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No. 777397 (iConsensus). The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA partners Bayer, Byondis, GSK, Pfizer, Rentschler, Sanofi, and UCB.

Nomenclature

MLPCA	Maximum Likelihood Principal Component Analysis
(N)MPC	(Nonlinear) Model predictive Control
$\mathrm{R_j}$	Set of reactants of reaction j
$\mathrm{P_j}$	Set of products of reaction j
AA	Amino acid
ξ	Vector of metabolite concentrations
n	Total number of metabolites (or species)
φ_j	Rate of the jth reaction
m	Total number of reactions
$\mathrm{k_{.,j}}$	Stoichiometric coefficient of element (.) in the jth reaction
K	Stoichiometric matrix
θ	Model parameter vector
θ_s	Stoichiometric parameter vector
θ_k	Kinetic parameter vector
ν	Transport vector
$\xi^{\Delta}_{\mathrm{f_i}}$	Differential transport-free state vector
$\hat{\rho}$	Affine subspace basis
$\hat{\mathrm{K}}$	Stoichiometric matrix estimate
G	Transformation matrix
$\mathrm{n_S}$	Number of measured samples
$\xi^{\Delta}_{\mathrm{f,meas_i}}$	Measured transport-free state vector
$\mathrm{Q_i}$	Covariance matrix of the measurement errors (noise)
$\xi^{\Delta,\mathrm{m}}_{\mathrm{f}}$	Maximum Likelihood estimate vector
$\mathrm{J_m}$	Maximum likelihood cost function
$\varphi^{\mathrm{j}}_{\mathrm{max}}$	Maximum rate parameter of the jth reaction
$\mathrm{h_M}$	Monod factor

$\mathrm{K}_{\xi_\mathrm{i},\mathrm{s}}$	Half-saturation parameter of the Monod factor activated by ξ_i
h_J	Jerusalimski factor
$\mathrm{K}_{\xi_\mathrm{i},\mathrm{inh}}$	Inhibition parameter of the Jerusalimski factor activated by ξ_i
J_lin	Linear cost function
J	Nonlinear cost function
Q_φ	Covariance matrix of the errors on the estimated rates
λ	Penalization parameter
p_k	Kinetic parameter vector dimension
$\varphi_{\theta_\mathrm{k}}$	Kinetic parameter sensitivity matrix
$\hat{\sigma}^2_\mathrm{k}$	A posteriori estimate of the rate measurement error variance
FIM_k	Fisher information matrix of the kinetic parameter estimation
$\hat{\mathrm{P}}_\mathrm{k}$	Estimate of the kinetic parameter estimation error covariance matrix
N_meas	Total number of measurements over one experiment
$\hat{\sigma}^2_{\mathrm{P}_\mathrm{k}}$	Estimate of the kinetic parameter estimation error variance
ϵ	Lower threshold of the parameter order of magnitude
η	Upper threshold of the parameter accuracy (identifiability)
θ_model	Model full parameter set (vector)
θ_IC	Model state variable initial conditions
J_ML	Full parameter set identification ML cost function
FIM	Fisher information matrix of the full parameter set estimation
$\hat{\sigma}^2$	Estimate of the full parameter set estimation error variance
$\hat{\mathrm{P}}$	Estimate of the full parameter set estimation error covariance matrix
n_θ	Full parameter set vector dimension
CI	Confidence intervals
H_p	Prediction horizon
H_c	Control horizon
H_r	Robust horizon
ϕ	Objective function
α	Weighting parameter
β	Weighting parameter
G_ref	Glucose tracking reference concentration
Gn_ref	Glutamine tracking reference concentration
ω_i	i^{th} Scenario weighting factor
F_aa	Amino acid medium feed rate
F_G	Glucose medium feed rate
F_Gn	Glutamine medium feed rate
g	Gram

L	Liter
d	Day
u	Input variable vector
t	Time
T_s	Sampling time
RMSE	Root Mean Square Error

References

[1] M. de Tremblay, M. Perrier, C. Chavarie, J. Archambault, Optimization of fed-batch culture of hybridoma cells using dynamic programming: single and multi-feed cases, Bioprocess. Biosyst. Eng. 7 (5) (1992) 229–234.

[2] S. Dhir, K. J. M. Jr., R. R. Rhinehart, T. Wiesne, Dynamic optimization of hybridoma growth in a fed-batch bioreactor, Biotechnology and Bioengineering 67 (2) (2000) 197–205.

[3] Z. Amribt, H. Niu, P. Bogaerts, Macroscopic modelling of overflow metabolism and model based optimization of hybridoma cell fed-batch cultures, Biochemical Engineering Journal 70 (2013) 196–209.

[4] R. H. D. Deken, The crabtree effect: A regulatory system in yeast, J . gen. Microbiol. 44 (1966) 149–156.

[5] B. Sonnleitner, O. Käppeli, Growth of *Saccharomyces cerevisiae* is controlled by its limited respiratory capacity : Formulation and verification of a hypothesis, Biotechnol. Bioeng. 28 (1986) 927–937.

[6] A. Provost, G. Bastin, Dynamic metabolic modelling under the balanced growth condition, Journal of Process Control 14 (7) (2004) 717–728.

[7] F. Zamorano, A. Vande Wouwer, R. M. Jungers, G. Bastin, Dynamic metabolic models of cho cell cultures through minimal sets of elementary flux modes, Journal of Biotechnology 164 (3) (2013) 409–422.

[8] S. Fernandes de Sousa, G. Bastin, M. Jolicoeur, A. Vande Wouwer, Dynamic metabolic flux analysis using a convex analysis approach: Application to hybridoma cell cultures in perfusion, Biotechnology and Bioengineering 113 (5) (2016) 1102–1112.

[9] E. Hagrot, H. Oddsdóttir, J. Gonzalez Hosta, E. W. Jacobsen, V. Chotteau, Poly-pathway model, a novel approach to simulate multiple metabolic states by reaction network-based model - application to amino acid depletion in cho cell culture, Journal of Biotechnology 259 (2017) 235–247.

[10] E. Hagrot, H. Oddsdóttir, M. Mäkinen, A. Forsgren, V. Chotteau, Novel column generation-based optimization approach for poly-pathway kinetic model applied to cho cell culture, Metabolic Engineering Communications 8 (2019) e00083.

[11] T. Abbate, S. Fernandes de Sousa, L. Dewasme, G. Bastin, A. Vande Wouwer, Inference of dynamic macroscopic models of cell metabolism based on elementary flux modes analysis, Biochemical Engineering Journal 151 (2019) 107325.

[12] T. Abbate, L. Dewasme, A. Vande Wouwer, P. Bogaerts, Adaptive flux variability analysis of hek cell cultures, Computers and Chemical Engineering 133 (2020) 106633.

[13] L. Hebing, T. Neymann, S. Engell, Application of dynamic metabolic flux analysis for process modeling: Robust flux estimation with regularization, confidence bounds, and selection of elementary modes, Biotechnology and Bioengineering 117 (7) (2020) 2058–2073.

[14] L. Dewasme, P. Bogaerts, A. Vande Wouwer, Monitoring of bioprocesses: mechanistic and data-driven approaches, Studies in Computational Intelligence, (Computational Intelligent Techniques for Bioprocess Modelling, Supervision and Control, Maria do Carmo Nicoletti, Lakhmi C. Jain, eds.), Springer Verlag, 2009, pp. 57–97.

[15] A. Oliveira, Biotechnology, big data and artificial intelligence, Biotechnol. J. 14 (8:e1800613).

[16] L. Dewasme, Neural network-based software sensors for the estimation of key components in brewery wastewater anaerobic digester: an experimental validation, Water Science and Technology 80 (10) (2020) 1975–1985.

[17] L. Dewasme, F. Cote, P. Filee, A. Hantson, A. Vande Wouwer, Macroscopic dynamic modeling of sequential batch cultures of hybridoma cells: An experimental validation, Bioengineering 4 (1) (2017) 17.

[18] O. Bernard, G. Bastin, On the estimation of the pseudo stoichiometric matrix for macroscopic mass balance modelling of biotechnological processes, Math. Biosci. 193 (2005) 51–77.

[19] J. Mailier, A. Donoso-Bravo, A. Vande Wouwer, On the derivation of simple dynamic models of anaerobic digestion using macroscopic bioreaction schemes, Mathematical and Computer Modelling of Dynamical Systems: Methods, Tools and Applications in Engineering and Related Sciences 19 (4) (2013) 301–318.

[20] A. Grosfils, A. Vande Wouwer, P. Bogaerts, Hybrid neural network models of bioprocesses: a comparative study, IFAC Proceedings Volumes 38 (1) (2005) 159–164.

[21] D. Lee, A. Jayaraman, J. S. Kwon, Development of a hybrid model for a partially known intracellular signaling pathway through correction term estimation and neural network modeling, PLoS Computational Biology 16 (12) (2020) e1008472.

[22] M. S. F. Bangi, J. S.-I. Kwon, Deep hybrid model-based predictive control with guarantees on domain of applicability, AIChE Journal (2023) e18012.

[23] M. Wang, R. Risuleo, E. Jacobsen, V. Chotteau, H. Hjalmarsson, Identification of nonlinear kinetics of macroscopic bio-reactions using multilinear gaussian processes, Computers and Chemical Engineering 133 (2020) 106671.

[24] M. Savageau, Biochemical systems analysis: I. some mathematical properties of the rate law for the component enzymatic reactions, Journal of Theoretical Biology 25 (3) (1969) 365–369.

[25] L. Dewasme, A. Richelle, P. Dehottay, P. Georges, M. Remy, P. Bogaerts, A. Vande Wouwer, Linear robust control of *S. cerevisiae* fed-batch cultures at different scales, Biochemical Engineering Journal 53 (1) (2010) 26–37.

[26] L. Dewasme, D. Coutinho, A. Vande Wouwer, Adaptive and robust linearizing control strategies for fed-batch cultures of microorganisms exhibiting overflow metabolism, Lecture Notes in Electrical Engineering 89 (2011) 283–305.

[27] L.-O. Santos, L. Dewasme, D. Coutinho, A. Vande Wouwer, Nonlinear model predictive control of fed-batch cultures of micro-organisms exhibiting overflow metabolism: Assessment and robustness, Computers and Chemical Engineering 39 (2012) 143–151.

[28] L. Dewasme, S. Fernandes, Z. Amribt, L. Santos, P. Bogaerts, A. Vande Wouwer, State estimation and predictive control of fed-batch cultures of hybridoma cells, Journal of Process Control 30 (2015) 50–57.

[29] F. Ibanez, P. A. Saa, L. Barzaga, M. A. DUarte-Mermoud, M. Fernandez-Fernandez, E. Agosin, J. Perez-Correa, Robust control of fed-batch high-cell density cultures: a simulation-based assessment, Computers and Chemical Engineering 155 (2021) 107545.

[30] M. Abadli, L. Dewasme, S. Tebbani, D. Dumur, A. Vande Wouwer, An experimental assessment of robust control and estimation of acetate concentration in escherichia coli bl21(de3) fed-batch cultures, Biochemical Engineering Journal 174 (2021) 108103.

[31] L. Hebing, F. Tran, H. Brandt, S. Engell, Robust optimizing control of fermentation processes based on a set of structurally different process models, Industrial and Engineering Chemistry Research 59 (6) (2020) 2566–2580.

[32] G. Bastin, D. Dochain, On-Line Estimation and Adaptive Control of Bioreactors, Vol. 1 of Process Measurement and Control, Elsevier, Amsterdam, 1990.

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

[34] N. Jerusalimski, N. Engamberdiev, Continuous cultivation of microorganisms, Vol. 517, Academic Press, New York, 1969.

[35] P. Bogaerts, J. Castillo, R. Hanus, A general mathematical modelling technique for bioprocesses in engineering applications, Syst. Anal. Model. Simul. 35 (1999) 87–113.

[36] R. Fekih-Salem, L. Dewasme, C. Cordeiro Castro, C. Nobre, A. Hantson, A. Vande Wouwer, Sensitivity analysis and reduction of a dynamic model of a bioproduction of fructo.oligosaccharides, Bioprocess and Biosystems Engineering 42 (2019) 1793–1808.

[37] S. Lucia, T. Finkler, S. Engell, Multi-stage nonlinear model predictive control applied to a semi-batch polymerization reactor under uncertainty, Journal of Process Control 23 (9) (2013) 1306 – 1319.

[38] A. Yousefi-Darani, O. Paquet-Durand, J. Hinrichs, B. Hitzmann, Parameter and state estimation of backers yeast cultivation with a gas sensor array and unscented kalman filter, Eng Life Sci. 21 (2021) 170–180.

[39] L. Ranzan, L. F. Trierweiler, B. Hitzmann, J. O. Trierweiler, Avoiding misleading predictions in fluorescence-based soft sensors using autoencoders, Chemometrics and Intelligent Laboratory Systems 223 (2022) 104527.

[40] L. Dewasme, A. Vande Wouwer, Experimental validation of a full-horizon interval observer applied to hybridoma cell cultures, International Journal of Control 93 (11) (2020) 2719–2728.