Title: Towards a Digital Twin: a hybrid data-driven and mechanistic digital shadow to forecast the evolution of lignocellulosic fermentations

Pau Cabaneros Lopez[1], Isuru Udugama[1], Sune Tjalfe Thomsen[2], Christian Roslander [3], Helena Junicke[1], Miguel Mauricio-Iglesias[4], Krist V. Gernaey[1]

[1]Process and Systems Engineering Center (PROSYS), Department of Chemical and Biochemical Engineering, Technical University of Denmark (DTU), Building 228A, 2800 Kgs. Lyngby, Denmark

[2]Department of Geosciences and Natural Resource Management, University of Copenhagen, Frederiksberg C, Denmark

[3]Department of Chemical Engineering, Lund University, Lund, Sweden.

[4]CRETUS Institute, Department of Chemical Engineering, Universidade de Santiago de Compostela, 15782, Santiago de Compostela, Spain.

Corresponding author: Krist V. Gernaey, kvg@kt.dtu.dk

Journal: **Biofuels, bioproducts and biorefining**

Abstract (max 250 words)

The high substrate variability and complexity of fermentation media derived from lignocellulosic feedstocks affect the concentration profiles and the length of the fermentations. Not accounting for such variability raises operational and scheduling issues and affects the overall performance of these processes. In this work, a hybrid soft sensor was developed to monitor and forecast the evolution of cellulose-to-ethanol fermentations. The soft sensor consisted of two modules (a data-driven and a kinetic model) connected sequentially. The data-driven module used a partial-least-squares model to estimate the current state of glucose from spectroscopic data. The kinetic model was recursively fit to the known concentrations of glucose to update the long-horizon predictions of glucose, xylose and ethanol. This combination of real-time data update from an actual fermentation process to a high fidelity kinetic model constitutes the basis of the digital twin concepts and allows for the better real-time understanding of complex inhibition phenomena caused by different inhibitors commonly found in lignocellulosic feedstocks. The soft sensor was experimentally validated with three different cellulose-to-ethanol fermentations and the results suggested that this method is suitable to monitor and forecast fermentations when the measurements provide reasonably good estimates of the real states of the system. These results would allow increasing the flexibility of the operation of cellulosic processes and adapting the scheduling to the inherent variability of such substrates.

1. Introduction

Using renewable feedstocks such as lignocellulosic waste to produce ethanol is a sustainable alternative to reduce the dependence on fossil fuels. However, the pretreatment of lignocellulosic biomass prior to the fermentation results in the formation of several byproducts with potent inhibitory effects on *Saccharomyces cerevisiae* [1,2]. Upon contact with these

inhibitors, *S. cerevisiae* undergoes a complex physiological response to compensate for their toxicity and to remove the inhibitors from the fermentation media [3]. This detoxification causes prolonged lag-phases during which *S. cerevisiae* utilizes most of its cellular resources to maintain internal homeostasis instead of using them for cell growth. This detoxification phase extends the fermentation time and reduces the productivity of the process. Furfural, 5-hydroxymethylfurfural, and acetic acid are the most common inhibitors derived from lignocellulosic feedstocks [1]. Their concentration can change notoriously between and within feedstocks due to the natural variability of the biomass caused by different weather conditions or agricultural practices. Such variability poses a challenge in cellulosic processes as the length of the lag-phase can change considerably depending on the concentration of inhibitors, reducing the productivity of the process and increasing the risk for contamination by lactic acid bacteria [4]. Not accounting for the substrate variability when operating cellulose-to-ethanol fermentation raises scheduling issues both, up-stream and down-stream of the fermentation and hinders its optimal operation at commercial scale. Therefore, methods able to account for the substrate variability and to predict the progress of the fermentations are necessary to improve the operation of such processes.

Kinetic models describing the correlation between the different state variables (e.g., glucose, ethanol, or biomass) from the initial conditions of each batch have been successfully used to predict the profile of non-cellulosic ethanol fermentations and to optimize their operations [5,6]. However, in the cellulose-to-ethanol processes, the complex effects of the different inhibitors challenge the development of the kinetic model [2]. Decades-long research on the effects of furfural or acetic acid on *S.cerevisiae* have shown that the inhibitors cause a system-wide response and affect not only the central carbon metabolism [7], but also the cell membrane composition [8], the internal pH [9] or the transcriptome of the cells [10]. This complex and not fully understood response is challenging to model mechanistically. Often, simple empirical terms

that modify the substrate uptake rates depending on the concentration of inhibitors are used to model the inhibitions. These terms rely on inhibition constants that need to be identified experimentally using parameter estimation. However, since these inhibition terms summarize complex phenomena into simple mathematical expressions, the validity of the models is often limited to conditions similar to the ones used to identify the models [11,12]. This poses a challenge in cellulose-to-ethanol fermentations where the high substrate variability results in significant changes in the composition of the media in between batches.

The objective of the current work is to develop and implement an iterative soft sensor that relies on the 'real-time' measurements of the glucose concentration to re-estimate the inhibition constants of a kinetic model to better describe the complex effects of the inhibitors. This approach increases the flexibility of the kinetic model by recursively adjusting the empirical terms to the dynamics of each fermentation, resulting in better long-horizon forecasts of the fermentation progress. A similar framework was proposed by Spann *et al.* [13] to monitor unmeasured variables in lactic acid bacteria fermentations from on-line measurements of pH and ammonia. Spann *et al.* [13] re-estimated the maximum growth rate and an empirical parameter that accounted for the length of the lag-phase of the fermentation, attaining better predictions of the system state variables. In the current work, the on-line glucose predictions were calculated using partial-least-squares (PLS) regression models from spectral data collected on-line using attenuated total reflectance mid-infrared spectroscopy (ATR-MIR). At each iteration, the inhibition constants were re-estimated and the new trajectories of the key state variables were plotted. This approach was implemented off-line and demonstrated experimentally using historical data from three cellulose-to-ethanol fermentations with different initial conditions simulating an on-line implementation. The proposed soft sensor corresponds to the recently emerged concept of a digital shadow of the process, i.e. a digital representation of a physical system that is then connected with the physical process in question in 'real-time' and is able to accurately represent the system [14–16]. This constitutes the basis for the

implementation of a digital twin, where the data does not only flow from the system to the model, but the model predictions can also actuate on the system to correct process deviations. The information flow between the fermentation system and the digital twin is illustrated in detail in **Figure 1**.

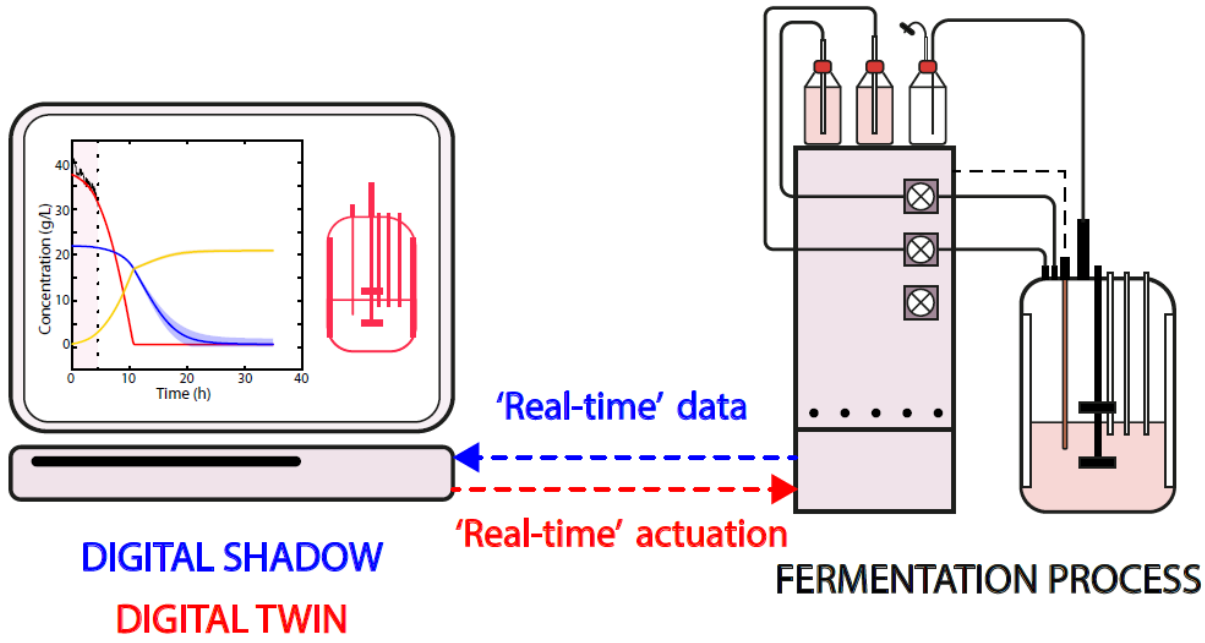

Figure 1. In a digital shadow, the data flow is unidirectional from the process to the digital model, while in a digital twin, the flow of information is bi-directional.

The current manuscript is organized as follows: Section 2 explains the experimental and numerical methods, section 3 describes the structure of the soft sensor, sections 4 and 5 show and discuss the results, and section 6 concludes the paper.

2. Materials and methods

2.1. Experimental methods

2.1.1. Media preparation

Danish wheat straw (provided by TK Energy ApS) was steam pretreated and enzymatically hydrolyzed before fermentation experiments took place. The pretreatment took place at the Department of Chemical Engineering at Lund University, Sweden. Five kg of dried wheat straw were soaked in water for 1 hour (at a mass ratio of 1:10) prior to filtration through a 1 mm filter for another hour. Then the wheat straw was press filtered for 5 minutes at 200 bars. The resulting straw (dry matter content of 39.98 %) was steam pretreated in seven batches of 1 kg at 200 ºC and 16.8 bar for 10 minutes. A total of 16.14 kg of pretreated wheat straw were generated and stored in containers in the freezer until use. The hydrolysis of the pretreated material was done at the Department of Geosciences and Natural Resource Management at the University of Copenhagen. The pretreated wheat straw was manually homogenized and the pH was adjusted to 5 using 5 M NaOH. The pretreated material was then hydrolyzed in 15 batches. In each batch, 1 kg of pretreated wheat straw was introduced into a 2 L bottle of low-density polyethylene (Kautex Textron, Bonn, Germany) and incubated with 5 mL of Cellic® CTec2 enzyme (corresponding to an enzyme/cellulose mass ratio of 6 %) at 50 ºC in a rotatory drum incubator (Termaks, Bergen, Norway). A total of 15 L of enzymatically hydrolyzed wheat straw was produced and stored at -20 ºC before use.

2.1.2. Fermentation experiments

The microorganism used in the fermentation experiments was the xylose-consuming *Saccharomyces cerevisiae* CEN.PK.XXX [17]. Prior to the fermentation, one colony of *S. cerevisiae* was transferred from an YPX agar plate (yeast extract (10 g/L, Microbiology Fermentech, Merk, New Jersey, USA), peptone (20 g/L, peptone from casein, Microbiology

Fermentech, Merk, New Jersey, USA) and xylose (20 g/L, Sigma Aldrich, Missouri, USA)) to a 250 mL shake flask containing 100 mL of liquid YPX media. After 36 hours of inoculation at 30 ºC and 180 rpm, one milliliter of grown cell culture was transferred to a 500 mL shake flask with 250 mL of YPX media. The cell culture grew at 30 ºC and 180 rpm before being inoculated in the fermenter. The dry weight of the cell culture was measured prior to the incubation in the fermenter following the protocol described in [18]. Three batch fermentations were performed in Sartorius BIOSTAT A® 2.5 L bioreactors (Sartorius, Göttingen, Germany) with a working volume of 1.5 L. The pH was controlled at 6.0 using 2 M NaOH and 2 M H_2SO_4, and the temperature was kept at 30 ºC using a heat jacket and a cooling finger. The fermenter was equipped with two Rushton impellers with 6 blades each and the stirring rate was kept at 450 rpm. Yeast extract (5 g/L) and peptone (10 g/L) were added to the wheat straw to supplement nitrogen, vitamins and trace elements. Prior to the fermentation, the wheat straw hydrolysate was centrifuged (Heraeuse Multifuge X3R, Thermo Scientific, Massachusetts, USA) at 4000 rpm for 20 min to reduce the concentration of suspended particles. Inocula of three different sizes were added to each fermentation to attain different fermentation profiles. All fermentations lasted between 25 and 35 hours and were stopped when all the xylose was consumed. An overview of the initial concentrations for the three batch fermentations is shown in **Table 1**.

3. **Table 1.** Summary of the initial concentrations for three different batch fermentations.

Initial concentrations	Fermentation 1	Fermentation 2	Fermentation 3
Size of inoculum (g/L)	1.0	0.4	1.4
Glucose (g/L)	37.0	42.0	39.0
Xylose (g/L)	22.0	22.0	22.0
Furfural (g/L)	0.6	0.6	0.6
Acetic acid (g/L)	3.0	3.0	3.3
5-HMF (g/L)	0.0	0.0	0.0

3.1. Analytical methods

3.1.1. HPLC analysis

Every hour, a fermentation sample was manually withdrawn for analysis with high-performance liquid chromatography (HPLC). The samples were immediately filtered through a 0.20 µm cellulose acetate filter (Labsolute, Renningen, Germany) and stored at -20 ºC. The off-line samples were analyzed with an Ultimate 3000 HPLC instrument equipped (Thermo Scientific, Massachusetts, USA) with 4 UV/VIS and a refractive index (RI) detector (ERC RefractoMax 520, Prague, Check Republic). An Aminex HPX-87 H column (BIORAD, California, USA) at 50 ºC with 5 mM H_2SO_4 as eluent was used to separate glucose, xylose, ethanol, furfural, acetic acid, and 5-HMF. The method lasted 80 minutes per sample, with an eluent flow rate of 0.6 mL/min. The samples were derivatized by adding 50 µL of H_2SO_4 to 950 µL of sample prior to the injection.

3.1.2. Spectroscopic analysis

ATR-MIR spectra (428-1833 cm^{-1} with a resolution of 1 cm^{-1}) were automatically collected on-line every minute using a spectrophotometer manufactured by NLIR ApS (Farum, Denmark) and provided by CellView IVS (Hillerød, Denmark). The instrument was equipped with a flow-cell connected to the fermenter using a closed recirculation loop with a flow rate of 90 mL/min. A background spectrum was measured with the laser turned off, and the reference spectrum was measured with the laser on and with air in the flow-cell. The exposure time of the sample was 120 milliseconds. Hundred spectra were collected every minute and their average was saved as an individual text file.

3.2. PLS calibration

The PLS regression model for glucose was calibrated by creating a set of synthetic samples that contained the fermentation matrix and uncorrelated concentrations of glucose, xylose and ethanol [19,20]. Xylose and ethanol were also accounted within the calibration due to the overlap of their spectra with glucose, and due to the fact that their dynamic profiles are correlated during the fermentation. Not including them as uncorrelated variables in the

calibration samples would result in interference of xylose and ethanol with the prediction of glucose. The fermentation matrix (without glucose, xylose or ethanol) was obtained using the following procedure. First, 1.5 L of wheat straw hydrolysate were fermented as described in **Section 2.1.2** to remove the glucose and the xylose from the media. Then, the biomass was removed by centrifuging the media at 4000 rpm for 10 minutes. Finally, the ethanol was stripped out from the media by sparging with sterile air at 35 ºC. After 24 hours, the volume was adjusted to 1.5 L using deionized water. The synthetic samples were thoroughly prepared to minimize the correlations between the concentrations of glucose, xylose and ethanol. The calibration ranges for glucose, xylose and ethanol were (0-40), (0-25) and (0-22) g/L, respectively. 100,000 Latin hypercube designs [21] with 21 samples each were randomly generated with different concentrations of glucose, xylose and ethanol. Then, the Pearson correlation coefficient (PCC) between glucose, xylose and ethanol was calculated for each design. The design with the lowest correlation was selected and applied experimentally (**Table 2)**.

Table 2. Selected Latin hypercube design. PCC Glucose/Xylose = -0.061, PCC Glucose/Ethanol = 0.002, PCC Xylose/Ethanol = 0.004.

Sample Nr.	Glucose (g/L)	Xylose (g/L)	Ethanol (g/L)*
0	2.00	1.25	12.00
1	0.00	17.50	7.50
2	34.00	5.00	21.00
3	38.00	21.25	28.50
4	36.00	16.25	9.00
5	24.00	6.25	18.00
6	18.00	15.00	16.50
7	12.00	25.00	13.50
8	26.00	2.50	22.50
9	22.00	12.50	19.50
10	10.00	10.00	6.00
11	8.00	23.75	30.00
12	4.00	3.75	24.00
13	30.00	7.50	4.50
14	32.00	0.00	0.00
15	20.00	18.75	10.50
16	28.00	13.75	25.50

17	40.00	22.50	3.00
18	14.00	11.25	15.00
19	16.00	8.75	27.00
20	6.00	20.00	1.50

* Note that the stripping process did not remove all the ethanol and a minimum concentration of 2.5 g/L was left in the media

Each sample was prepared by using 75 mL of fermentation matrix and by spiking the corresponding amount of glucose, xylose, and ethanol. The spectrum of each sample was collected as described in **Section 2.2.2**. All the data analysis was done using Python 3.7. First, the spectrum of each sample was preprocessed by taking its first derivative, filtering them using a Savitzky-Golay filter and mean centering it. Then, a PLS1 model for glucose was calibrated using the MBPLS library [22]. An optimal number of 4 latent variables was found by minimizing the root-mean-square error of cross-validation (RMSECV) during a leave one out cross-validation routine (LOO-CV). The resulting model described more than 95 % of the variance in the spectral matrix and in the concentrations vector.

3.3. Kinetic model

A dynamic model of the fermentation was used to forecast the evolution of the state variables of the fermentation. The model was based on the model developed by Mauricio-Iglesias *et al.* [23]. It described the growth of *S. cerevisiae* on glucose and xylose and accounted for the inhibition effects of furfural and acetic acid. 5-HMF was not considered because it was not present in the fermentation media. The different processes considered in the model are listed in **Table 3**.

Table 3. List of processes considered in the kinetic model.

Processes considered in the kinetic model
1. Uptake of glucose
2. Uptake of xylose
3. Uptake of furfural

4. Furfural is detoxified into furfuryl alcohol
5. Furfuryl alcohol inhibits the uptake of glucose
6. Furfuryl alcohol inhibits the uptake of xylose
7. Furfural inhibits the uptake of glucose
8. Furfural inhibits the uptake of xylose
9. Uptake of acetic acid
10. Acetic acid inhibits the uptake of glucose
11. Acetic acid inhibits the uptake of xylose
12. Production of ethanol
13. Ethanol inhibits the uptake of glucose
14. Ethanol inhibits the uptake of xylose
15. Cell growth
16. Glucose inhibits the uptake of xylose

The mass balance of the state variables was described using the stoichiometric matrix (**Table 4**).

Table 4. Stoichiometric matrix explaining the mass balance of the system. Glucose (Glu), Xylose (Xyl), Furfural (Fur), Furfuryl alcohol (FA), Acetic acid (HAc), Ethanol (EtOH) and Biomass (X).

	Glucose	**Xylose**	**Furfural**	**FA**	**HAc**	**Ethanol**	**Biomass**
Glucose Uptake	-1	0	0	0	0	$Y_{EtOH/Glu}$	$Y_{X/Glu}$
Xylose Uptake	0	-1	0	0	0	$Y_{EtOH/Xyl}$	$Y_{X/Xyl}$
Furfural Uptake	0	0	-1	$Y_{FA/Fur}$	0	0	0
HAc Uptake	0	0	0	0	-1	0	0

The uptake rates of the substrates, glucose and xylose, were described using Monod type kinetics with substrate inhibition (**Eq. 1** and **2**), while the uptake rates of the inhibitors (furfural and acetic acid) were described as simple Monod kinetics (**Eq. 3** and **4**). Empirical expressions were multiplied by the uptake rates of glucose and xylose to account for the inhibitory effects of furfural, furfuryl alcohol, acetic acid, ethanol, and catabolite repression (in the uptake rate of xylose) (**Eq. 1** and **2**).

Eq. 1
$$v_{glu} = X \cdot \frac{v_{max,Glu} \cdot Glu}{K_{SP,Glu} + Glu + \frac{Glu^2}{K_{i,Glu}}} \cdot \left(1 - \left(\frac{EtOH}{P_{max,Glu}}\right)^{\gamma_{Glu}}\right) \cdot \left(\frac{K_{i,Fur,Glu}}{1 + Fuyr}\right) \cdot \left(\frac{1}{1 + \frac{FA}{K_{i,FA,Glu}}}\right) \cdot \left(\frac{1}{1 + \frac{HAc}{K_{i,HAc,Glu}}}\right)$$

Eq. 2
$$v_{Xyl} = X \cdot \frac{v_{max,Xyl} \cdot Xyl}{K_{SP,Xyl} + Xyl + \frac{Xyl^2}{K_{i,Xyl}}} \cdot \left(1 - \left(\frac{EtOH}{P_{max,Xyl}}\right)^{\gamma_{Xyl}}\right) \cdot \left(\frac{1}{1 + \frac{Fur}{K_{i,Fur,Xyl}}}\right) \cdot \left(\frac{1}{1 + \frac{FA}{K_{i,FA,Xyl}}}\right) \cdot \left(\frac{1}{1 + \frac{HAc}{K_{i,HAc,Xyl}}}\right) \cdot \left(\frac{1}{1 + \frac{Glu}{K_{i,Glu,Xyl}}}\right)$$

Eq. 3
$$v_{Fur} = X \cdot \frac{v_{max,Fur} \cdot Fur}{K_{SP,Fur} + Fur}$$

Eq. 4
$$v_{HAc} = X \cdot \frac{v_{max,HAc} \cdot HAc}{K_{SP,HAc} + HAc}$$

The resulting model, consisting of 7 ordinary differential equations with 26 parameters (**Supplementary materials**), was implemented in Matlab 2016® (Mathworks, USA). First, the model was solved using the parameters found by Mauricio-Iglesias *et al.* but it did not describe the experimental data used in this work (the fermentations were much faster than what the model predicted). This was probably because instead of using ammonia salts, the fermentation media used in this work was supplemented with yeast extract and peptone as source of vitamins, minerals and nitrogen, which increase the maximum biomass specific uptake. Therefore, in order to get a reasonable description of the experimental data, the uptake rates for glucose, xylose and acetic acid ($v_{max,Glu}$, $v_{max,Xyl}$ and $v_{max,HAc}$) were re-estimated using parameter estimation. The concentration profile of glucose, xylose, ethanol, furfural, and acetic acid of a historical fermentation were used to fit the kinetic model by applying the non-linear least-squares method using the *lsqnonlin* function of Matlab 2016® (data not shown). The parameters found by Mauricio-Iglesias *et al.* [23] were used as an initial guess for the parameter estimation, and the initial conditions were the same as in fermentation 3 but with 23 g/L of xylose (**Table 1**). After re-estimating the specific growth rates, the model had a reasonable good fit with the

experimental data (**Figure 4**). Once the model was identified, a local sensitivity and identifiability analysis were done to find which inhibition constants could be iteratively re-estimated in 'real-time' in the soft sensor. The sensitivity revealed which inhibition constants had a higher impact on the model outputs, and the identifiability analysis showed subsets of parameters that are approximately linearly independent. A detailed explanation of the local sensitivity and identifiability analysis is given by Sin and Gernaey [24] and in Brun *et al.* [25]. Subsets of parameters with a collinearity index below 20 were considered to be approximately linearly independent [25]. These results were used to determine the most relevant subset of inhibition constants to be re-estimated on-line using the soft sensor.

4. Monitoring framework

The structure of the proposed soft sensor is shown in **Figure 2**. It is an iterative algorithm consisting of two sequential modules: a data-driven model followed by a kinetic model that updates the forecast of the state variables (concentration of glucose, xylose, and ethanol) of the fermentation every 15 minutes. The data-driven module uses a PLS regression model to estimate the current concentration of glucose from the collected ATR-MIR spectra. Then, the kinetic model is recursively fit to the measured concentration of glucose using least-squares minimization to update the kinetic parameters of the model. The updated kinetic model is used to forecast the progress of the fermentation. The uncertainty in the parameters is also calculated and propagated through the kinetic model using Monte Carlo simulations. To evaluate the performance of the soft sensor, the root-mean squared error (RMSE) between the model predictions and the off-line samples was calculated at each iteration (**Eq. 5**)

Eq. 5

$$RMSE_n = \sqrt{\frac{\sum_{k=1}^{K}(x - \hat{x})^2}{K}}$$

Where n is the iteration number, K is the total number of off-line samples measured, x is the value of the state variables measured with HPLC and $\hat{x}$ is the predicted value of the state variables.

4.1. Module 1: the data-driven model

The objective of the first module is to provide an estimate of the current state of glucose at a given time t_n. With that objective in mind, spectra of the fermentation media are collected and individually stored as separate .txt files every minute during the fermentation. Every 15 minutes, all the spectra collected between t_0 and time t_n, are imported into a matrix M, which becomes the input of the data-driven module. Before predicting the concentration of glucose, all spectra are preprocessed by taking their first derivative, filtering using Savitzky-Golay and finally mean centering. Then, a calibrated PLS1 regression model is used to calculate the glucose

concentration from the pretreated spectra. The output of the data-driven module is a vector ($\hat{y}$) containing all the predicted glucose concentrations from t_0 to t_n.

4.2. Module 2: the kinetic model

Module 2 is comprised of two steps: a model updating followed by a forecasting step. In the model updating step, the empirical parameters of the kinetic model are updated by fitting the kinetic model to the estimated glucose concentration from t_0 to t_n. Then, the forecasting step uses the updated kinetic model to forecast the progress of the fermentation from t_n to t_{end}.

4.2.1. Model updating step

At t_n, the model updating step takes the vector ($\hat{y}$) with the predicted glucose concentration between t_0 and t_n as input. The parameters corresponding to the inhibition terms of furfuryl alcohol and acetic acid on the glucose uptake ($K_{i,FA,Glu}$ and $K_{i,HAc,Glu}$) are estimated by fitting the kinetic model to the vector ($\hat{y}$) using maximum likelihood estimation (MLE). The search space for each parameter was limited to ±2.5 % of its value at time t_n during each iteration to increase the robustness of the parameter estimation. Only the inhibition constants of furfuryl alcohol and acetic acid on glucose (and not on xylose) were re-estimated because xylose was only consumed after glucose depletion due to the strong catabolite repression of glucose on xylose. The uncertainty associated with the estimated parameters was assessed using the bootstrap method [24]. In brief, a set of 50 synthetic data sets were created by randomly sampling from the residual distribution of the model identified using MLE. Then, the two parameters were re-estimated using MLE for each of the 50 synthetic data sets resulting in 50 estimates of each parameter. Finally, the mean, the standard deviation, and the covariance matrix of the two parameters were calculated. The newly estimated parameters were used as the initial guess of the parameters for the next iteration.

319

320 *4.2.2. Forecasting step*

321 The forecasting step takes as input the newly identified parameters and the initial conditions of
322 the process to make long-horizon forecasts (from t_n to t_{end}) of the trajectories of the key state
323 variables (glucose, xylose, and ethanol) of the fermentation. The uncertainty associated with the
324 parameter was propagated through the model using the Monte Carlo method [24]. The
325 parameter estimate covariance matrix represented the uncertainty of the estimated parameters
326 ($K_{i,FA,Glu}$ and $K_{i,HAc,Glu}$), while the remaining model parameters were considered to be
327 independent and normally distributed with a standard deviation corresponding to 1 % of their
328 mean values. A total of 100 Monte Carlo samples were calculated at each iteration resulting in a
329 population of 100 forecasts. Then, the mean values and 95 % CI were calculated from the
330 resulting population.

331

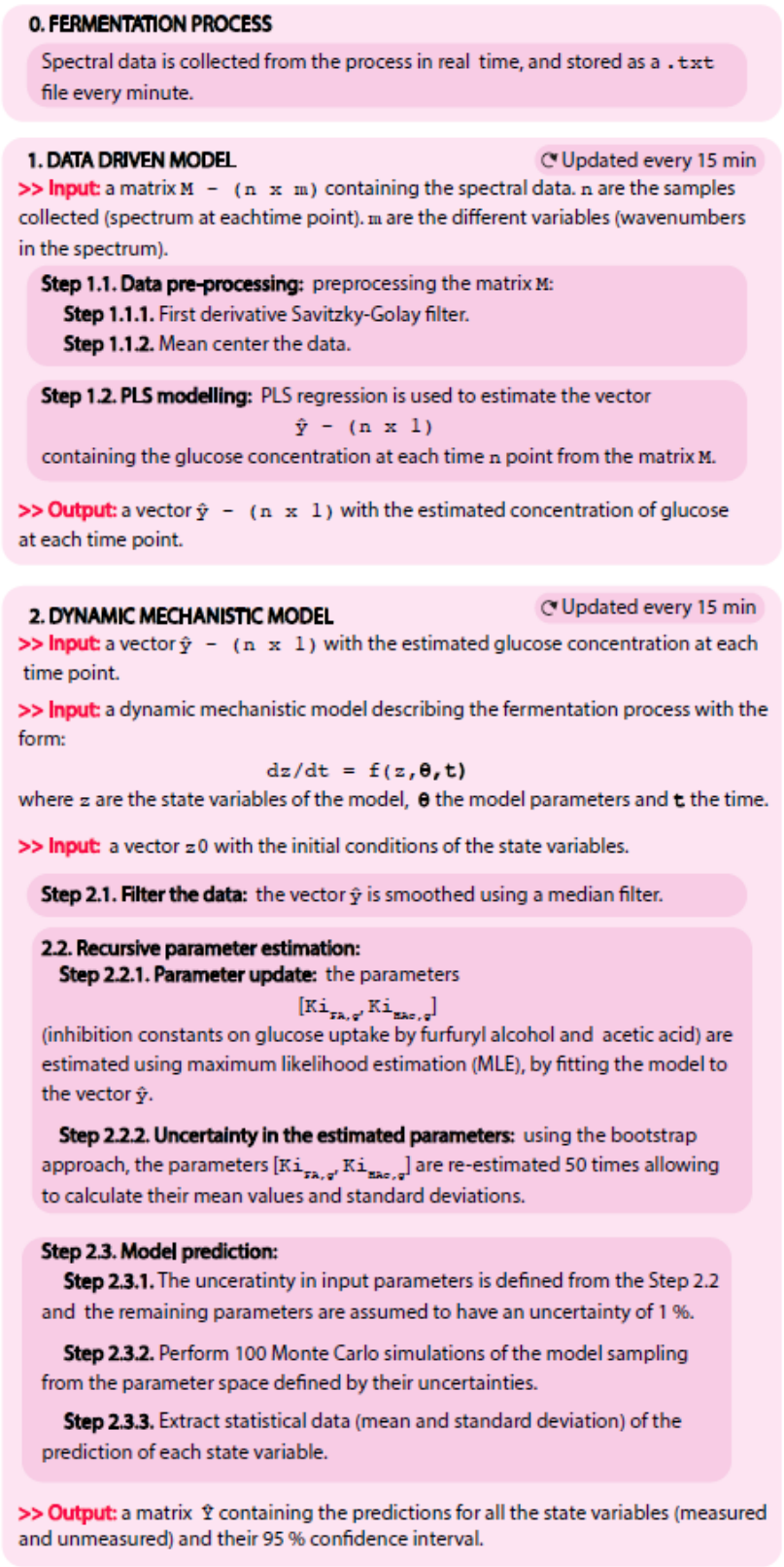

Figure 2. Schematic representation of the algorithm of the proposed soft sensor.

5. Results

5.1. On-line glucose prediction using PLS models

The PLS regression model was calibrated using synthetic samples in order to decouple the concentrations of glucose, xylose, and ethanol. The calibrated PLS model successfully predicted the glucose concentration in the calibration set with an RMSECV of 1.45 g/L (**Figure 3.A**). Since no fermentation samples were included in the calibration set, the PLS model was directly validated using fermentations 1-3. In fermentation 1, the glucose concentration was accurately predicted by the PLS model (**Figure 3.B**), having an RMSE of 2.55 g/L. In fermentations 2-3, the PLS model described the general evolution in the glucose concentration, but in both fermentations the models overestimated the glucose concentration compared to the off-line reference measurements. This lack of fit was reflected in the RMSE of the predictions (7.66 g/L and 8.30 g/L for fermentations 2 and 3 respectively, **Figure 3.B)**. In general, the PLS predictions in fermentation 2 had a higher bias than in fermentation 3, but a clog in the sampling line during fermentation 3 stopped the flow through the flow-cell for 30 minutes and caused deviating PLS predictions, which resulted in a higher RMSE. This situation is illustrated in **Figure 3.B**, where most of the PLS prediction in fermentation 3 is closer to the real value than for fermentation 2, and can be seen in **Figures 5 B.1-5** and **C.1-5**.

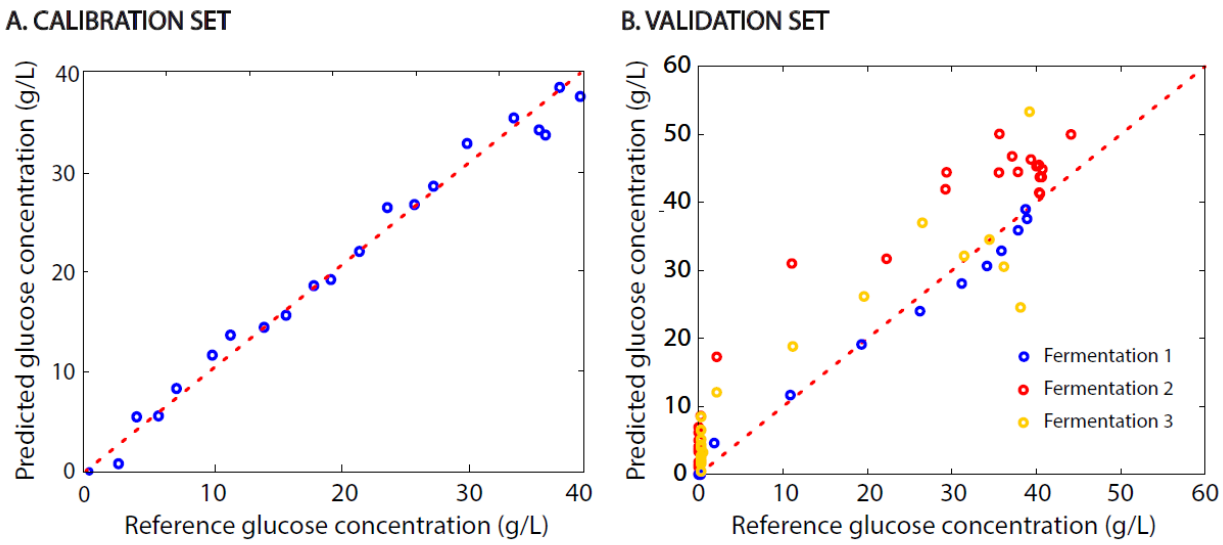

Figure 3. A. Predicted versus measured glucose concentration in the calibration set. The root-mean-square error during leave one out cross-validation (RMSECV) was 1.45 g/L. **B.** Predicted versus measured glucose concentration in the validation fermentations 1-3. The root-mean-square error was 2.55 g/L, 7.66 g/L and 8.30 g/L for fermentations 1, 2, and 3, respectively.

5.2. Kinetic model: parameter estimation and sensitivity analysis

The results of the parameter estimation showed a general satisfactory fit with the experimental data (**Figure 4**). The model was able to describe the concentration profiles of glucose, ethanol and acetic acid accurately (**Figures 4.A, 4.C,** and **4.E**). Xylose was predicted reasonably well, but the model did not capture the consumption of xylose towards the end of the fermentation (**Figure 4.B**). The uptake of furfural was only described well at the beginning of the fermentation, **Figure 4.D**).

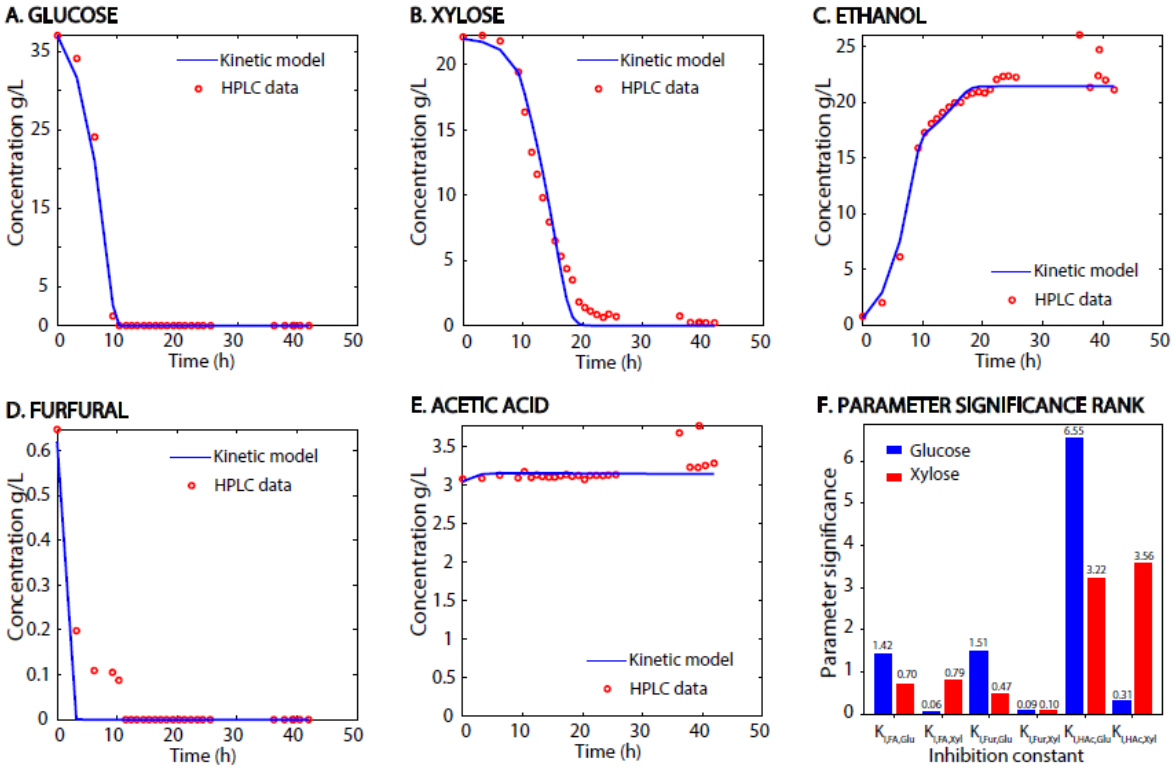

Figure 4. Results of the parameter estimation. **Figures 4.A-E** show the results of the fit for glucose (Glu), xylose (Xyl), acetic acid (HAc), furfural (Fur), and ethanol. **Figure 4.F** shows the results of the local sensitivity analysis, shown as the parameter significance of the different inhibition constants on the profile of glucose and xylose. $K_{i,FA,Glu}$ and $K_{i,FA,Xyl}$ are the inhibition constants of furfuryl alcohol (FA) on glucose and xylose respectively, $K_{i,Fur,Glu}$ and $K_{i,Fur,Xyl}$ are the inhibition constants of furfural on glucose and xylose respectively and $K_{i,HAc,Glu}$ and $K_{i,HAc,Xyl}$ are the inhibition constants of acetic acid on glucose and xylose.

Once the model was identified, a local sensitivity and identifiability analysis were performed first to study the effect of the inhibitory constants on the output of the model, and then to find a subset of constants that could be simultaneously identified in the on-line implementation of the soft sensor. The local sensitivity analysis revealed important insights about the dynamics of the model (**Figure 4.F**). First, it showed that while glucose was only sensitive to the inhibition constants that directly affected its uptake (i.e., $K_{i,FA,Glu}$, $K_{i,Fur,Glu}$, and $K_{i,HAc,Glu}$), xylose was sensitive to all the inhibition constants that affected the uptake of glucose and xylose (i.e.,

$K_{i,FA,Glu}$, $K_{i,Fur,Glu}$, $K_{i,HAc,Glu}$, $K_{i,FA,Xyl}$, $K_{i,Fur,Xyl}$, and $K_{i,HAc,Xyl}$). This was due to the catabolite repression that glucose exerts on xylose (i.e., glucose inhibited the uptake of xylose). This result agrees with the profiles obtained experimentally, showing that furfural, glucose, and xylose were consumed sequentially. The sensitivity analysis also revealed that acetic acid was the inhibitor with a larger impact on the concentration of both, glucose and xylose (**Figure 4.F**). This is because acetic acid was not consumed and its inhibitory effects remained throughout the fermentation. A similar situation happened with furfural and furfuryl alcohol. Although furfuryl alcohol is a much weaker inhibitor than furfural (the inhibition constants for glucose are 5.00 and 0.75 g/L for furfuryl alcohol and furfural, respectively), the direct inhibition of furfuryl alcohol on glucose ($K_{i,FA,Glu}$) affected glucose and xylose almost as much as $K_{i,Fur,Glu}$. This was also because furfuryl alcohol accumulated during the fermentation, while furfural was rapidly consumed within the first three hours (**Figure 4.D**). Because of furfural, glucose and xylose were consumed sequentially; xylose was not directly affected by furfural and it was only indirectly affected through $K_{i,FA,Glu}$. A subset containing all the inhibition constants would ideally be selected to be estimated in the on-line implementation of the soft sensor. However, the identifiability analysis revealed that this was not possible due to the high degree of correlation between the different parameters. Among the uncorrelated subsets of parameters (with a collinearity index below 25), the $K_{i,FA,Glu}$, and $K_{i,HAc,Glu}$ were selected to be estimated because they had a significant effect on both, the concentration of glucose and xylose.

5.3. On-line implementation of the soft sensor

The soft sensor was implemented simulating an on-line system using data of three batch fermentations. Every 15 minutes, the soft sensor took the PLS measurements to re-identify the kinetic model and to make long-term predictions of the trajectories of glucose, xylose, and ethanol.

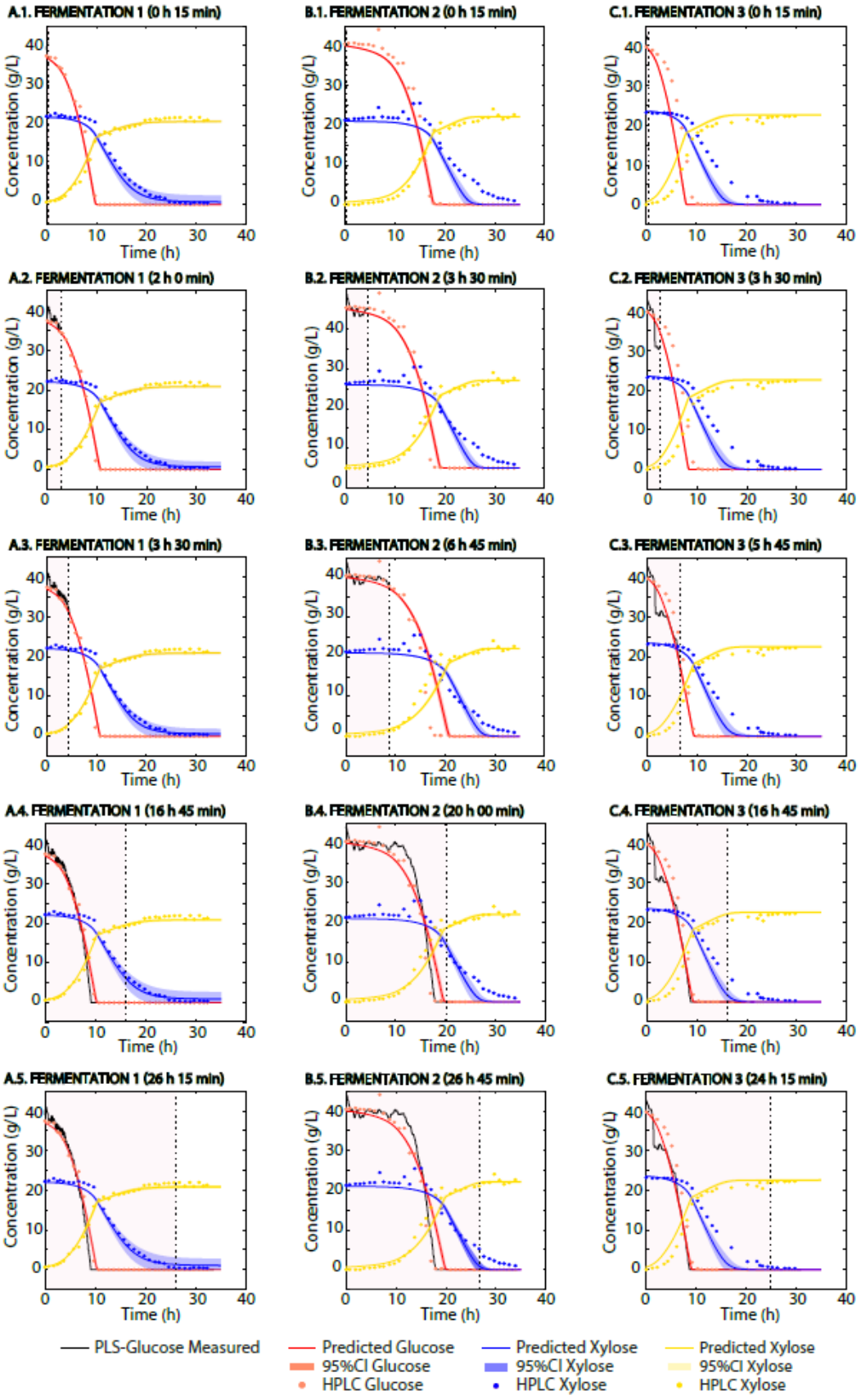
A.1. FERMENTATION 1 (0 h 15 min)
B.1. FERMENTATION 2 (0 h 15 min)
C.1. FERMENTATION 3 (0 h 15 min)
A.2. FERMENTATION 1 (2 h 0 min)
B.2. FERMENTATION 2 (3 h 30 min)
C.2. FERMENTATION 3 (3 h 30 min)
A.3. FERMENTATION 1 (3 h 30 min)
B.3. FERMENTATION 2 (6 h 45 min)
C.3. FERMENTATION 3 (5 h 45 min)
A.4. FERMENTATION 1 (16 h 45 min)
B.4. FERMENTATION 2 (20 h 00 min)
C.4. FERMENTATION 3 (16 h 45 min)
A.5. FERMENTATION 1 (26 h 15 min)
B.5. FERMENTATION 2 (26 h 45 min)
C.5. FERMENTATION 3 (24 h 15 min)
Concentration (g/L)
Time (h)
PLS-Glucose Measured
Predicted Glucose
95%CI Glucose
HPLC Glucose
Predicted Xylose
95%CI Xylose
HPLC Xylose
Predicted Xylose
95%CI Xylose
HPLC Xylose

Figure 5. State variables predicted with the soft sensor at five different time points during the fermentation. **Figures A.1-5** show the predictions made for fermentation 1, **Figures B.1-5** for fermentation 2, and **Figures C.1-5** for fermentation 3. The dashed line indicates the time at which the prediction was made. The PLS measurements before the dashed line were used to re-identify the kinetic model allowing to make long-horizon predictions of the evolution of the fermentation.

Figure 5 shows the predicted trajectories of glucose, xylose, and ethanol made by the kinetic model at different time points during the three batch fermentations. Since the predictions of the model were updated at each iteration, the RMSE of the predictions with the off-line measured data was calculated at each iteration to evaluate the performance of the soft sensor (**Figure 6**).

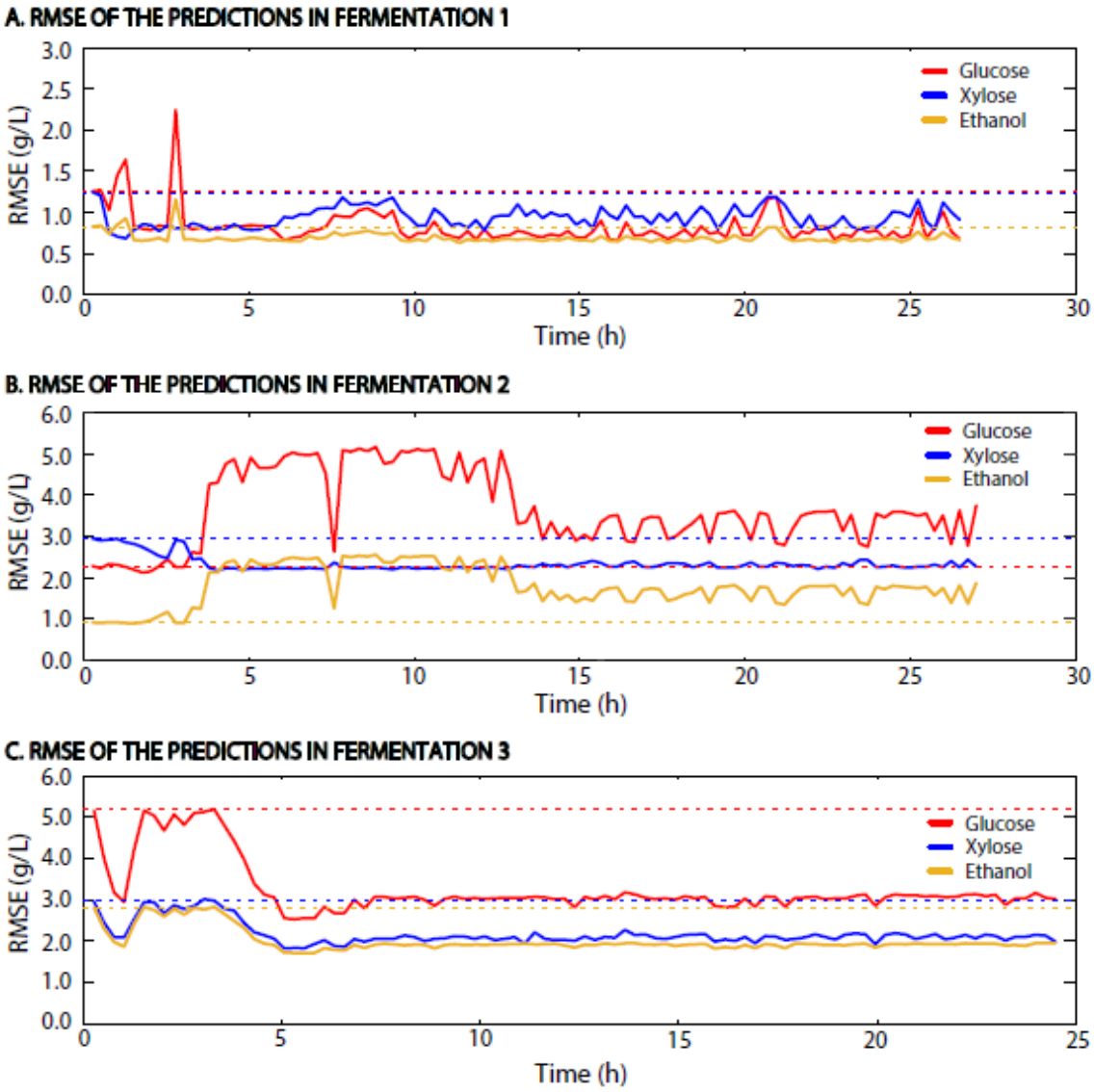

Figure 6. Root-mean square errors (RMSE) of the kinetic model predictions at each iteration of the soft sensor. **A**. Fermentation 1. **B.** Fermentation 2. **C.** Fermentation 3. The dashed lines in plots **A-C** indicate the RMSE of the model at time 0.

By updating the kinetic model at each iteration, the soft sensor improved the long-term predictions of the kinetic model in the course of fermentations 1 and 3 (**Figures 5.A.1-5** and **C.1-5**), resulting in a considerable reduction of the RMSE for all three variables (**Figures 6.A** and **C**). After 2-5 hours of fermentation, it was possible to obtain a good and stable prediction of the progress of glucose, xylose and ethanol. However, this improvement was not found in

fermentation 2 (**Figure 5.B**), where the predictions of the soft sensor were less stable and only improved the kinetic model in the prediction of xylose. This is likely due to the biased prediction of the glucose concentration made by the PLS model (**Figure 3.B**). In all fermentations, the RMSE of the prediction became worst during the glucose consumption phase, where the measured glucose concentration changed quickly between two consecutive iterations.

6. Discussion and perspectives for application

The hybrid soft sensor presented in this work iteratively uses the glucose concentration (measured on-line with a PLS model) to re-estimate the empirical parameters of a kinetic model and to obtain better long-horizon predictions of the concentrations of glucose, xylose, and ethanol. This procedure allowed updating the kinetic model at each iteration, providing it with a higher degree of flexibility to describe the inhibitory effects, and accounting for the actual development of the fermentation. This method was conceptualized to monitor and forecast complex fermentation processes with high substrate variability, where kinetic models would fail at predicting the evolution of the fermentation (e.g., cellulose-to-ethanol fermentations). The validation experiments showed that the soft sensor outperformed the kinetic model in two out of three batch fermentations, where the soft sensor provided stable predictions of the evolution of glucose, xylose and ethanol (with a low RMSE) after 2-5 hours of fermentation (fermentations 1 and 3, **Figure 6.A** and **C,** respectively). However, the soft sensor did not perform better than the kinetic model in one of the validation experiments (fermentation 2, **Figure 6.B**). The poor prediction of the soft sensor in fermentation 2 was likely caused by the higher bias of the PLS measurements in that experiment (**Figure 3.B**). This shows a limitation of the proposed monitoring algorithm. Since the kinetic model was updated at each iteration, it was assumed that the PLS measurements represented a good estimator of the real system state. This is not necessarily true as most 'real-time' monitoring methods of fermentation processes tend to be noisy (especially those based on vibrational spectroscopy [26]). In fermentations 1 and 3, the soft sensor performed well because the PLS measurements of the glucose concentration were reasonably close to the real value. However, in fermentation 2, PLS prediction errors were confounded with the effect of the inhibitors and incorporated within the parameters of the kinetic model, resulting in deviating predictions. In this work, the PLS model was calibrated using synthetic samples designed to minimize the correlation between the concentrations of glucose,

xylose, and ethanol, and to account for the matrix of the system. However, the changes in the fermentation matrix were not considered within the calibration set, and these could have interfered with the prediction of the glucose concentration. Improving the calibration set (for instance, by including real fermentation samples to account for the changes in the matrix) would potentially improve the predictions of the PLS model. However, even when the PLS models are calibrated adequately, the measurements may not represent reasonable estimates of the system. In these cases, the estimation error is then transmitted to the values of the estimated parameters and, ultimately, to the model predictions. An alternative approach would be to consider the information from the PLS models not as a perfect measurement of the state variables, but as an estimate of each of the states. In this case, state estimators (e.g., Kalman filters) can be used to evaluate the most likely state of the system combining the information provided by the PLS and the kinetic model [27,28].

7. Conclusions

The soft sensor proposed in this work iteratively used a data-driven model and a kinetic model to make long-horizon predictions of the evolution of different state variables during cellulose-to-ethanol fermentations. At each iteration, the data-driven model was used to estimate the current state of glucose from spectroscopic measurements. Then, the kinetic model was fit to the measured glucose concentration and used to make long-horizon predictions of the evolution of the fermentation. This approach adds flexibility to the kinetic model allowing it to adapt the predictions to the known development of the fermentation and it is appropriate to monitor processes where kinetic models use simple empirical terms to describe complex phenomena, and it constitute a simple digital twin demonstration. The experimental results showed that the soft sensor performed better than the kinetic model only, when the PLS glucose concentration predictions were close to the real glucose concentration. This denotes the dependence of the soft sensor on the quality of the monitoring method, as it assumes that the measurements provide a reasonable estimate of the real state of the process. This assumption may limit the applicability of the proposed soft sensor to processes, where the measuring method provides reliable estimates of the state variables. In processes, where the measurements are not reliable enough, it would be interesting to use state estimators such as variants of the Kalman filter for non-linear systems.

8. List of abbreviations

ATR-MIR	Attenuated total refractance mid infrared spectroscopy
PLS	Partial least squares
HPLC	High-performance liquid chromatography
Glu	Glucose
Xyl	Xylose
Fur	Furfural
FA	Furfuryl alcohol
5-HMF	5-hydroxymethyl furfural
HAc	Acetic acid
EtOH	Ethanol
X	Yeast biomass

9. Acknowledgements

The authors wish to thank Associate Prof. Carl Johan Franzén from Chalmers University of Technology, Sweden for providing the xylose-consuming yeast strain, *Saccharomyces cerevisiae* CEN.PK.XXX. This project has been partially supported by the Energy Technology Development and Demonstration Program (EUDP) in the frame of the project “Demonstration of 2G ethanol production in full scale” (Grant number 64015-0642), and has been realized together with Maabjerg Energy Center (MEC) and Novozymes A/S. The authors wish to acknowledge the support provided by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement number 713683 (COFUNDfellowsDTU), by the Danish Council for Independent Research in the frame of the DFF FTP research project GREENLOGIC (grant agreement number 7017-00175A), and by Novo Nordisk Fonden in the frame of the Fermentation-Based Biomanufacturing education initiative. MIM belongs to the Galician Competitive Research Group ED431C 2017/029 and the CRETUS Strategic Partnership (AGRUP2017/01).

525

526

527

Bibliography

[1] E. Palmqvist, B. Hahn-Hägerdal, Fermentation of lignocellulosic hydrolysates . I : Inhibition and detoxification, (2000) 17–24. doi:10.1016/S0960-8524(99)00160-1.

[2] E. Palmqvist, B. Hahn-Hägerdal, Fermentation of lignocellulosic hydrolysates. II: Inhibitors and mechanisms of inhibition, Bioresour. Technol. 74 (2000) 25–33. doi:10.1016/S0960-8524(99)00161-3.

[3] E. Palmqvist, H. Grage, N.Q. Meinander, B. Hahn-Hägerdal, Main and interaction effects of acetic acid, furfural, and p- hydroxybenzoic acid on growth and ethanol productivity of yeasts, Biotechnol. Bioeng. 63 (1999) 46–55. doi:10.1002/(SICI)1097-0290(19990405)63:1<46::AID-BIT5>3.0.CO;2-J.

[4] C. Li, J.E. Aston, J.A. Lacey, V.S. Thompson, D.N. Thompson, Impact of feedstock quality and variation on biochemical and thermochemical conversion, Renew. Sustain. Energy Rev. 65 (2016) 525–536. doi:10.1016/j.rser.2016.06.063.

[5] J.H.T. Luong, Kinetics of Ethanol Inhibition in Alcohol Ferment at ion, Biotechnol. Bioeng. 27 (1985) 280–285.

[6] A. Nanba, Y. Nishizawa, Y. Tsuchiya, S. Nagai, Kinetic analysis for batch ethanol fermentation of Saccharomyces cerevisiae, J. Ferment. Technol. 65 (1987) 277–283. doi:10.1016/0385-6380(87)90088-4.

[7] E. Palmqvist, J. Almeida, B. Hahn-Hagerdal, Influence of furfural on anerobic glycolytic kinetics of Saccharomyces cerevisiae in batch culture., Biotechnol. Bioeng. 62 (1999) 447–454.

[8] L. Lindberg, A.X.S. Santos, H. Riezman, L. Olsson, M. Bettiga, Lipidomic Profiling of Saccharomyces cerevisiae and Zygosaccharomyces bailii Reveals Critical Changes in Lipid Composition in Response to Acetic Acid Stress, PLoS One. 8 (2013) 1–12. doi:10.1371/journal.pone.0073936.

[9] M. Palma, J.F. Guerreiro, I. Sá-Correia, Adaptive response and tolerance to acetic acid in Saccharomyces cerevisiae and Zygosaccharomyces bailii: A physiological genomics perspective, Front. Microbiol. 9 (2018) 1–16. doi:10.3389/fmicb.2018.00274.

[10] G. Wu, Z. Xu, L.J. Jönsson, Profiling of Saccharomyces cerevisiae transcription factors for engineering the resistance of yeast to lignocellulose-derived inhibitors in biomass conversion, Microb. Cell Fact. 16 (2017). doi:10.1186/s12934-017-0811-9.

[11] A. Golabgir, C. Herwig, Combining Mechanistic Modeling and Raman Spectroscopy for Real-Time Monitoring of Fed-Batch Penicillin Production, Chemie-Ingenieur-Technik. 88 (2016) 764–776. doi:10.1002/cite.201500101.

[12] D. Dochain, Bioprocess Control, 2010. doi:10.1002/9780470611128.

[13] R. Spann, C. Roca, D. Kold, A. Eliasson Lantz, K. V. Gernaey, G. Sin, A probabilistic model-based soft sensor to monitor lactic acid bacteria fermentations, Biochem. Eng. J. 135 (2018) 49–60. doi:10.1016/j.bej.2018.03.016.

[14] W. Kritzinger, M. Karner, G. Traar, J. Henjes, W. Sihn, Digital Twin in manufacturing: A categorical literature review and classification, IFAC-PapersOnLine. 51 (2018) 1016–1022. doi:10.1016/j.ifacol.2018.08.474.

[15] E.H. Glaessgen, D.S. Stargel, The Digital Twin Paradigm for Future NASA and U.S. Air Force Vehicles, in: 53rd Struct. Struct. Dyn. Mater. Conf. Spec. Sess. Digit. Twin, 2011.

[16] R. He, G. Chen, C. Dong, S. Sun, X. Shen, Data-driven digital twin technology for optimized control in process systems, ISA Trans. 95 (2019) 221–234. doi:10.1016/j.isatra.2019.05.011.

[17] J.O. Westman, N. Bonander, M.J. Taherzadeh, C.J. Franzén, Improved sugar co-utilisation by encapsulation of a recombinant Saccharomyces cerevisiae strain in alginate-chitosan capsules, Biotechnol. Biofuels. 7 (2014) 1–14. doi:10.1186/1754-6834-7-102.

[18] E. El-Mansi, C. Bryce, B. Hartley, A. Demain, Fermentation Microbiology and

Biotechnology, in: Ferment. Microbiol. Biotechnol. Third Ed., 2012: pp. 47–98. doi:10.1201/b11490-2.

[19] N. Petersen, P. Ödman, A.E. Cervera Padrell, S. Stocks, A.E. Lantz, K. V. Gernaey, In situ near infrared spectroscopy for analyte-specific monitoring of glucose and ammonium in Streptomyces coelicolor fermentations, Biotechnol. Prog. 26 (2010) 263–271. doi:10.1002/btpr.288.

[20] K. Pontius, Monitoring of Bioprocesses Opportunities and Challenges Monitoring of Bioprocesses Opportunities and Challenges (PhD Thesis), Technical University of Denmark, 2019. https://orbit.dtu.dk/en/publications/monitoring-of-bioprocesses-opportunities-and-challenges-opportuni.

[21] D.C. Montgomery, Design and Analysis of Experiments, 8th ed., 2009.

[22] A. Baum, L. Vermue, Multiblock PLS: Block dependent prediction modeling for Python, J. Open Source Softw. 4 (2019) 1190. doi:10.21105/joss.01190.

[23] M. Mauricio-Iglesias, K. V. Gernaey, J.K. Huusom, State Estimation in Fermentation of Lignocellulosic Ethanol. Focus on the Use of Ph Measurements, in: 12th Int. Symp. Process Syst. Eng. 25th Eur. Symp. Comput. Aided Process Eng. Copenhagen, Denmark., Copenhagen, Denmark, 2015.

[24] G. Sin, K. V. Gernaey, Data Handling and Parameter Estimation, in: Exp. Methods Wastewater Treat., IWA Publishing, London, UK, London, UK, 2016: pp. 201–234.

[25] R. Brun, P. Reichert, Practical identifiability analysis of large environmental simulation, 37 (2001) 1015–1030.

[26] P. Cabaneros, H. Feldman, M. Mauricio-iglesias, H. Junicke, J. Kjøbsted, K. V Gernaey, Benchmarking real-time monitoring strategies for ethanol production from lignocellulosic biomass, Biomass and Bioenergy. 127 (2019) 105296. doi:10.1016/j.biombioe.2019.105296.

[27] D. Krämer, R. King, A hybrid approach for bioprocess state estimation using NIR

spectroscopy and a sigma-point Kalman filter, J. Process Control. 82 (2017) 91–104. doi:10.1016/j.jprocont.2017.11.008.

[28] D. Krämer, R. King, On-line monitoring of substrates and biomass using near-infrared spectroscopy and model-based state estimation for enzyme production by S. cerevisiae, IFAC-PapersOnLine. 49 (2016) 609–614. doi:10.1016/j.ifacol.2016.07.235.

[29] M.S. Krishnan, N.W.Y. Ho, G.T. Tsao, Fermentation kinetics of ethanol production from glucose and xylose by recombinant Saccharomyces 1400(pLNH33), Appl. Biochem. Biotechnol. - Part A Enzym. Eng. Biotechnol. 77–79 (1999) 373–388. doi:10.1385/abab:78:1-3:373.

[30] T.J. Hanly, M. a. Henson, Dynamic model-based analysis of furfural and HMF detoxification by pure and mixed batch cultures of S. cerevisiae and S. stipitis, Biotechnol. Bioeng. 111 (2014) 272–284. doi:10.1002/bit.25101.