Massimo Morbidelli ORCID iD: 0000-0002-0112-414X

Hybrid-EKF: Hybrid Model coupled with Extended Kalman Filter for real-time monitoring and control of mammalian cell culture

Authors: Harini Narayanan[1], Lars Behle[1], Martin F. Luna[1], Michael Sokolov[1,2], Gonzalo Guillén-Gosálbez[1], Massimo Morbidelli[2,3], Alessandro Butté[2] *

[1]Institute of Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich, Switzerland

[2]DataHow AG, Zurich, Switzerland

[3]Dipartimento di Chimica, Materiali e Ingegneria Chimica, Giulio Natta, Politecnico di Milano

Corresponding Author: Alessandro Butté, DataHow AG, Zurich, Switzerland

E-mail: a.butte@datahow.ch

Abstract

In a decade when industry 4.0 and Quality by Design are major technology drivers of biopharma, automated and adaptive process monitoring and control are inevitable requirements and model-based solutions are key enablers in fulfilling these goals. Despite strong advancement in process digitalization, in most cases, the generated

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/bit.27437.

data sets are not sufficient for relying on purely data-driven methods, while the underlying complex bioprocesses are still not completely understood. In this regard, hybrid models are emerging as a timely pragmatic solution to synergistically combine available process data and mechanistic understanding. In this work we show a novel application of Hybrid-EKF framework, that is hybrid models coupled with extended Kalman filter for real-time monitoring, control and automated decision making in mammalian cell culture processing. We show that, in the considered application, the predictive monitoring accuracy of such framework improves by at least 35% when developed with hybrid models with respect to industrial benchmark tools based on PLS models. Additionally, we also highlight the advantages of this approach in industrial applications related to conditional process feeding and process monitoring. With regards to the latter, for an industrial use case we demonstrate that the application of Hybrid-EKF as a soft sensor for titer shows a 50% improvement in prediction accuracy compared to state-of-the-art soft sensor tools.

Keywords: Bioprocessing, Hybrid Models, Extended Kalman Filter, Process Monitoring, Adaptive control.

1 Introduction

The emergence of new digital technologies and artificial intelligence has set the ground for the process industries to step towards industry 4.0. One of the key pillars of this revolution is the establishment of smart factories that can adaptively manage a variety of process scenarios and operate autonomously, i.e. without human intervention (Catlin, Lorenz, Sternfels, & Willmott, 2017). This requires robust

monitoring and control of processes in real-time (Narayanan et al., 2020). However, in biopharmaceutical industries a characteristic difficulty is to quantify in real-time critical quality attributes (CQAs) of the product (Zhang, 2009). This imposes the need for indirect methods, such as soft sensors, to successfully implement the Process Analytical Technology (PAT) goals in an Industry 4.0 format (Flickinger et al., 2013; Narayanan et al., 2020; Teixeira, Oliveira, Alves, & Carrondo, 2009).

Soft sensors are mathematical models supporting hardware sensors to estimate process- or product-relevant variables (Mandenius & Gustavsson, 2015). Spectroscopic techniques such as Raman and NIR coupled with calibration models based on multivariate (Mehdizadeh et al., 2015) or advanced machine learning tools (Voss, Mittelheuser, Lemke, & Luttmann, 2017) are particularly attractive for such application. A single probe typically can provide information about multiple process variables such as cell density (Ohadi, Legge, & Budman, 2015a), metabolite and amino acid concentration (Berry, Moretto, Matthews, Smelko, & Wiltberger, 2015; Bhatia, Mehdizadeh, Drapeau, & Yoon, 2017) and CQAs (Li, Ray, Leister, & Ryder, 2013). However, in order to be able to use these estimations directly for monitoring and control applications, the calibration models must be sufficiently accurate and reliably cover a significantly large process operation range (Solle et al., 2017). In contrast, currently the signal acquisition from online devices in bioprocesses are noisy and can be subject to additional offsets due to non-optimal calibration models.

Kalman Filter (KF) is a filtering technique that uses a linear mathematical model (Kalman, 1960) in combination with real-time sampling from the dynamic system to produce improved state estimations compared to measurements or model predictions

alone. Extended (Jazwinski, 1970) and Unscented (Simon, 2006) Kalman Filter (E/U)KF are variations of the original formulation which were adapted for non-linear systems. There are countless examples of the KF in the chemical industry (Mohd Ali, Ha Hoang, Hussain, & Dochain, 2015) where the mathematical models are well established. However, this is not the case for bioprocesses due to our limited mechanistic understanding. Additionally, central process variables are measured at different and (or) not sufficiently large frequencies. This had initially hampered the application of the KF to bioprocesses, but now the situation is changing with more powerful analytical and digitalization technologies, as well as modeling approaches.

Most implementations of (U/E) KF are reported for microbial cultures wherein first principle (FP) based mathematical models are coupled with measurements, in the simple case of pH and dissolved oxygen sensors (Dewasme, Goffaux, Hantson, & Wouwer, 2013; Rimvydas Simutis & Lübbert, 2017; Svrcek, Elliott, & Zajic, 1974; Wang, Zhao, & Yu, 2010), or of more complex ones such as those based on Raman and NIR spectroscopy (Feidl et al., 2019; Krämer & King, 2017; Markana, Padhiyar, & Moudgalya, 2018). Additionally, application of EKF together with fuzzy assisted first principle models (R. Simutis, Havlik, & Lübbert, 1992) and artificial neural networks (R. Simutis, Havlik, & Lübbert, 1993) for control of fermentation reactors have also been demonstrated. However, only a few examples exist for the use of EKFs in mammalian culture (Ohadi et al., 2015a) but also using FP based models.

It is worth noting that, for all the aforementioned works with FP models, experiments are performed specifically to fit the parameters of the model. This constrains the possibility of using historical data for such models due to parameter identifiability

restrictions (Franceschini & Macchietto, 2008). Hybrid modeling is emerging as an attractive approach in this direction as it combines the advantages of FP and data-driven models (Narayanan, Sokolov, Morbidelli, & Butté, 2019; von Stosch et al., 2014) and enables learning from historical data. The combination of a hybrid model with EKF has only been demonstrated for *Saccharomyces Cerevisiae* (Zorzetto & Wilson, 1996). While, for mammalian cells the potential application of hybrid models for monitoring and model predictive control has been advocated in a couple of cases in the past (Dors, Simutis, & Lübbert, 1995; Gnoth, Jenzsch, Simutis, & Lübbert, 2008; Lübbert & Simutis, 1994; Sommeregger et al., 2017; von Stosch et al., 2014). However, no concrete application has been presented to this end, especially using hybrid model as an alternative to the mechanistic model in the EKF for mammalian cell cultures.

In this work, we demonstrate a modeling framework combining hybrid model with EKF (referred to as Hybrid-EKF) for mammalian cell culture applications and highlight the advantage of such a framework in monitoring and control. The framework is developed on a *simulated* dataset which provides the true values of the states ideally required to verify the hypothesis of better state-estimation by EKF and additionally allows to vary measurement frequency and noise levels of relevant variables to, for instance, mimic spectroscopic acquisitions. A feed control application and a soft sensor application for titer estimation on real bioreactor data are highlighted. Further, the superiority of the developed Hybrid-EKF algorithm to industrial benchmark tools is quantified.

2 Materials and methods

2.1 Dataset

The Hybrid-EKF framework developed in this work was first tested on a *simulated cell culture process dataset*. Once the framework development protocol is established, same procedure is applied to a real industrial cell culture dataset.

2.1.1 Data organization

In line with the concept that has been introduced in our previous work on Hybrid models (Narayanan, Sokolov, Morbidelli, et al., 2019), the process data collected are grouped into different information sources as represented schematically in Figure 1A. The operating conditions that stay constant throughout the process are denoted as two-dimensional matrix, Z, with rows and columns representing runs and operating conditions, respectively. The dynamically changing, non-controlled process variables such as viable cell density (X_v) and Glucose (GLC) are represented as three-dimensional matrix X, which features an additional time dimension in comparison to the Z matrix. Both the simulated and real dataset are obtained from a fed-batch system with bolus feed addition after metabolite measurement. The mass of different metabolites added during the culture time is organized into a different three-dimensional matrix referred as F. Another dynamic information source, included in the W matrix, are the variables which are controlled in time across a fixed set-point such as pH, temperature (T), pCO_2 and pO_2. Finally, the product characteristics are denoted by matrix Y. The variables constituting X and Y matrix were used as the states to represent the dynamic evolution of the system. The

variables constituting the different information matrices are tabulated in Figure 1B for both the datasets.

2.1.2 In-silico Dataset

The dataset was simulated from the macro-kinetics models for mammalian cell cultures proposed by Craven, Shirsat, Whelan, & Glennon, 2013 and Xing, Bishop, Leister, & Li, 2010),with adaptations to impose complex non-linearities in growth rate dependencies and to account for pH and temperature shifts. The *in-silico* model consists of 14 process factors, namely initial conditions of X_v, GLC, GLN, pH before and after shift, temperature before and after shift, days of shift, glucose and glutamine feeding start day-end day and the amount of bolus feed for glucose and glutamine respectively (kept constant for all the days of feeding). A fractional factorial design was used to simulate 100 experiments (supplementary file) with each experiment running for 14 days. 100 experiments were simulated to give a fair chance to the data-driven models to perform well and to be comparable the same number of experiments were used to develop the hybrid model. However, robust hybrid models with comparable accuracy in prediction can be obtained with about 15 experiments as shown in SI Figure 1.

Given the nomenclature defined in Figure 1, the process variables (X) include, viable cell density (X_v), concentration of glucose (GLC), lactate (LAC), glutamine (GLN), and ammonium (NH_4) and the process target (Y) was titer. Measurements were simulated at a frequency of 2.4 h with 15% added Gaussian noise. Simulated data are used to mimic measurements from a spectroscopic technique. Thus, an error to represent spectroscopic soft sensor accuracy is chosen (Feidl et al., 2019; Ohadi,

Legge, & Budman, 2015b). On the other hand, pH and temperature were considered in matrix W, as there were planned dynamic shifts performed during the experiment. Additionally, GLC and GLN were fed on a daily basis, whereby the feed starting and end times were also amongst the designed factors. There was no Z information in the simulated dataset, as all of the designed factors could be represented in the other information matrices.

2.1.3 Real Dataset

The cell culture process dataset originally published and described in detail by (Rouiller et al., 2012) was used to further assess the performance of the framework. The dataset contains 81 fed-batch runs (3.5 L working volume) with each experiment running for 10 days. Experiments were performed by manipulating three seed train conditions namely, the N-1 amplification process cell density, duration, and cell age as well as two process conditions, that is the set points of pH and dissolved oxygen (DO), which constitute the Z matrix. The X matrix was constituted by X_v, GLC, LAC, GLN, glutamate (GLU), NH_4 and osmolality (Osm). While, pH, pO_2 and pCO_2 constitutes the W matrix and titer is denoted in Y. The X and W information was measured once per day, while Y is measured on even days from day 0 until the end of the run. The preprocessing steps such as interpolation and missing data imputation of the dataset are adopted from (Narayanan, Sokolov, Morbidelli, et al., 2019).

2.2 Methodology

The dataset is divided into three parts: calibration set for model development, tuning set to determine EKF parameters (c.f. section 2.2.2) and an independent test set to assess the model performance. The split for the two used datasets is demonstrated in

Figure 2A. During model calibration, cross validation strategy is used to tune the model hyper parameters (Hastie, Tibshirani, & Friedman, 2017). The EKF tuning, required for the calculation of the Process Noise covariance matrix, is performed using the tuning set experiments (detailed description in section 2.2.2). For comparison purposes, the evaluation of the accuracy and performances of model and model-EKF framework are made on the basis of Root Mean Squared Error in Prediction (RMSEP). Data-driven models such as ANNs has previously been used as benchmark to show superior performance of hybrid models (Feyo de Azevedo, Dahm, & Oliveira, 1997; Von Stosch, Oliveira, Peres, & Feyo De Azevedo, 2012). However, in this work Historical-PLS2 model, which is the industrial benchmark, is used for comparison. Two variants, namely, Historical PLS2 models coupled with Extended Kalman Filter (PLS-EKF) and the classical Historical PLS2 with the actual measurements (PLS-direct) is used. All computations are performed using MATLAB R2018b.

2.2.1 Models

Two different types of models are coupled with the EKF, hybrid model and Historical-PLS2 model. A detailed description of hybrid modeling procedure used in this work is presented in (Narayanan, Sokolov, Morbidelli, et al., 2019). The system of equation representing the cell culture is established based on mass balances, as follows:

$$\frac{dState_i}{dt} = \mu_i(t)\, X_v(t) \qquad (1)$$

where $State_i$ and μ_i are the concentrations and specific rates of the *i–th* species, respectively, and *i* represents X_v, GLC, GLN, LAC, NH_4 and titer for *in-silico* dataset

and X_v, GLC, LAC, GLN, GLU, NH_4, Osm and titer for the real bioreactor dataset. Artificial Neural Networks with $States_i$, W and Z as the input are used to compensate for the lack of deterministic understanding of the specific rates (μ_i). The number of nodes in the single hidden layer (with L-2 regularization) is tuned using five-fold cross-validation that resulted in 8 and 10 nodes for *in-silico* and real dataset, respectively.

Two Historical-PLS2 based models, namely PLS-EKF and PLS-direct, were used for benchmark comparison. The detailed description of Historical-PLS model has been reported in (Narayanan, Sokolov, Butté, & Morbidelli, 2019). Essentially, a PLS2 model per time point is constructed to map all the available data until that time to the states of the six considered variables at the immediate next time. As a result, the Historical-PLS2 model has an independent model, one per time point. The mapping is represented as shown in the equation below:

$$[Z, X(0 < t \leq t_{model}), W(0 < t \leq t_{model})] \xrightarrow{PLS2} State(t_{model+1}) \quad (2)$$

where t_{model} denotes the time point at which the PLS2 model is developed and the States are nothing but X and Y variables (as mentioned in section 2.2.1). However, it should be noted that during the prediction phase, for PLS-propagated, only input at t = 0 (i.e. the process design) is given to the Historical-PLS2 model. Thereby, the value at time t=1 is predicted by the model as shown in *equation 2*, which then serves as an input to the model for predicting the next time point t = 2, and so on and so forth. On the other hand, for PLS-direct, a classical Historical-PLS2 model (Sokolov et al., 2017) was developed, where the actual measurement is used at time t = 1 (instead of the prediction from previous model or corrected state) to predict states at

time t = 2, so on and so forth. In all cases, Monte-Carlo-based five-fold cross-validation is used to determine the optimal number of latent variables. All the PLS models were developed using the in-built MATLAB function *plsregress ()* which is based on the SIMPLS algorithm (Tie Jong, 1993).

2.2.2 Extended Kalman Filter (EKF)

The main idea of using the EKF is to produce better estimates of known or unknown variables by combining measurements with statistical noise to predictions made by a model of the system. The mathematics of the EKF has been well established and can be found in literature such as (Jazwinski, 1970; Schiff, 2012; Simon, 2006). In a nutshell, EKF takes two noisy information sources that is the model prediction (*equation 3*) with noise $q_{(t)}$ and the actual measurement with noise $r_{(t_k)}$ (*equation 4*) and provides an estimation that is better than either of the information source, when tuned appropriately.

$$\frac{d(State)}{dt} = f(State(t), W(t), Z) + q(t) \quad (3)$$

$$State^{meas} = h\big(State\ (t_k)\big) + r(t_k) \quad (4)$$

Here, f is the dynamic model (Hybrid or PLS-propagated in our case) representing the states of the system while h is the measurement function that indicates if a certain state variable is measured or not. Since *in-silico* dataset was used to mimic a measurement from spectroscopic sensor, all the state variables were assumed to be measurable. Whereas for the real dataset all variables except titer were considered as measured to highlight the ability of the EKF to also improve the estimate of unmeasured states. The noise terms q and r are assumed to be Gaussian with zero

mean and covariance matrixes Q and R, respectively. The combination of the two information sources (measurement and model prediction) at the discrete sampling point t_k is achieved using equation 5 and represented schematically in Figure 2B:

$$State^{Corr}(t_k) = State^{Pred}(t_k) + K(t_k)(State^{meas}(t_k) - h(State^{Pred}(t_k)))$$ (5)

where matrix K is called the gain of filter calculated directly by the EKF algorithm. At the sampling point t_k, a prediction is made with the model, referred to as the predicted states, by using the corrected states of t_{k-1} as input to the model. Corrected estimates of the state variables at t_k (referred to as corrected states) are then obtained by combining the measurements obtained at t_k with the predicted states. The figure legends and super indices, Pred and Corr refer to predicted and corrected states, respectively.

Lastly, in order to apply the EKF algorithm, the covariance matrices (Q and R) related to the two noise terms (q and r as per equation 3 and 4) have to be specified. The latter is usually defined as a diagonal matrix containing the variance of the measurement method for every measured variable (which is defined by the hardware sensor and the calibration method). On the other hand, the calculation of Q is one of the most challenging parts of the implementation (Schneider & Georgakis, 2013). Here, Q is considered as a diagonal matrix with the values of the model's Mean Squared Error (MSE) per variable in training, multiplied by a tuning parameter k_Q:

$$Q\ (t) = \ k_Q\ diag(MSE_i\ (t))$$ (6)

An optimization problem that minimizes the MSE of the corrected states estimated by the model-EKF with respect to the true values of the tuning dataset is posed in order to fit k_Q using MATLAB in-built function *patternsearch()*. On the other hand, for the industrial use case MSE of the corrected titer estimates with respect to measured titer on the tuning set is minimized to optimize k_Q. The predicted and corrected states when k_Q is optimized is correspondingly denoted as Opt-Pred and Opt-Corr.

3 Results

3.1 Protocol Validation

Firstly, to emphasis on our tuning procedure, a comparison of the performance of Hybrid-EKF framework using just the pragmatic guess of process noise (that is $k_Q = 1$) and the effect of multiplying with k_Q is made. Figure 3 presents the time resolved RMSEP of the four key variables, namely, X_v, Titer, GLC and LAC computed with respect to the true values of these states, i.e., the ones generated with the *in-silico* model without noise sampled at discrete time points (i.e., every 2.4 h). The blue lines (Meas) represent the noise in measurement, while the red line (Hyb) represents the error of the hybrid model in prediction. It can be observed for the case of 15% perturbed noise, the RMSEP of the measurement and model are quite comparable for all the variables except for glucose, where the model is slightly better than the measurements. The yellow line (Corr (Hyb)) shows the RMSEP of Hybrid-EKF framework using the pragmatic estimate of process and measurement noise covariance (that is when k_Q = 1) whereas the purple line (Opt-Corr (Hyb)) highlights the RMSEP of the framework when the process noise is optimized based on k_Q on a

tuning set. It can be observed that the Hybrid-EKF framework with the non-optimized noise (in yellow) is already better that the measurement or the model alone. However, it is still close to one of the two information sources. On contrary, once the process noise covariance is optimized (in purple), the performance of the Hybrid-EKF framework surpasses other sources of state estimation. For instance, for the estimation of final day titer, an improvement of 57.3%, 71.4% and 46% in the accuracy can be observed as compared to the measurement, model and non-optimized Hybrid-EKF framework, respectively. The RMSEP of the Hybrid–EKF framework with respect to the true values of the states (available from the *in-silico* model) averaged over all the runs and time points are 0.18×10^6 Cells/mL, 5.9 mM, 4.95 mM and 30.56 mg/L which is highly accurate compared to the range of the variables being 8×10^6 Cells/mL, 470 mM, 230 mM and 1700 mg/L, respectively. This shows that the EKF, when tuned appropriately, is much closer to the actual state of the system than either the model or the measurement alone, which confirms the theoretical basis of these filters. The value of k_Q obtained after optimization is 0.049 indicating that the pragmatic approach overestimates the process noise. Additionally, it is also worth highlighting that the defined framework is quite stable throughout the run, denoted by the non-growing errors for all for variables (comparing purple with other evolutions in Figure 3). Additional simulation studies for validating the protocol at different noise levels are presented in the supplementary information. This study shall be considered as validation of the promising framework development protocol, which will be used for the following analyses.

3.2 Hybrid-EKF comparison with benchmark tools

The Hybrid-EKF framework is now compared with the equivalent approach developed using the PLS-propagated instead of Hybrid as the base model. Using the same protocol (i.e. $k_Q$tuning), a PLS-EKF module is developed. Additionally, the classical Historical-PLS2 model denoted as PLS-direct (c.f. 2.2.1) is also used as a benchmark comparison. Considering first the base models, it can be observed in Figure 4A for an exemplary run in the test set, that the PLS-propagated model (purple line) used in a completely predictive manner predicts the true X_v trajectory (in black) very poorly. Subsequently, during tuning, the EKF trusts only the measurement neglecting the model. This observation is validated by Figure 4A which demonstrates a comparison of the X_v profile estimated by the PLS-EKF module (Opt-Corr (PLS), grey) against the classical PLS2 model (PLS-direct, crimson). On the other hand, in Figure 4B it can be observed that hybrid models (red dashed) predicts the trajectories of the variables closer to the actual values (represented in black). The state estimations are further improved by combining hybrid models with the EKF as highlighted by the green line in Figure 4B and C. The comparison of the Hybrid-EKF estimation with the PLS-direct model and the PLS-EKF estimation is illustrated with the profile of X_v for an exemplary run in Figure 4B and C, respectively. It can be observed that both the PLS direct and Opt-Corr (PLS) show large fluctuations and offset with respect to the true trajectory whereas the Opt-Corr (Hyb) fluctuates much less across the true value. A more detailed discussion of the trend is provided in supplementary information (c.f. SI section 2). In addition, Figure 4D then compares the RMSEP of the estimations averaged over all times and runs made by the hybrid model and three other modules, namely, the PLS-direct, PLS-

EKF and Hybrid-EKF for the four key state variables. It demonstrates that the Hybrid-EKF outperforms the reference tools by at least 35%, 32%, 34% and 25% for X_v, Titer, GLC and LAC, correspondingly. Thus, it can be concluded that selection of appropriate modeling approach and tuning procedure is essential for the efficient implementation of EKF. It has been demonstrated that Hybrid model coupled with an optimized EKF framework can outperform standard tools and be more accurate than even the hybrid models. Such a modeling approach is then a suitable option for monitoring, adaptive control and real-time automatic decision-making applications. This point is further explored in the next section of the article.

3.3 Monitoring and Feeding of Glucose using Hybrid-EKF

After having showcased the advantages of the Hybrid-EKF method, we next investigate the ability of the developed framework to support process control through model-based glucose monitoring and feeding. In Figure 5A it is visible that both the predicted and corrected states of the Hybrid-EKF fluctuates closer to the true value of the GLC with a comparable RMSEP of 5.9 mM. This is much less than the noise in the measurement or hybrid model alone which produces a RMSE of 16.0 mM and 9.3 mM, respectively. The accurate estimation from the predicted and corrected states facilitates the use of this framework for feed-forward and feed-back control, respectively. To demonstrate the added value of this framework for glucose control, an experiment is simulated with a conditional feeding of GLC such that, when the concentration of GLC in the system reaches below 20 mM, feeding is triggered to obtain a concentration after feeding equal to 30 mM. Figures 5B and C, show in black the profile of GLC if the true values of the states were accessible and

correspondingly the actual feed triggers that would have occurred in the ideal case are shown as black line in Figure 5D. Now this ideal case is compared to two real scenarios, namely to control glucose feeding based on measurements only (in yellow) and based on the state estimation from Hybrid-EKF framework (in green). Due to the accurate state estimation of GLC by the Hybrid-EKF the feed points are close to the ideal scenario, as shown in Figure 5D leading to a final GLC profile well in agreement with the ideal profile (Figure 5B). However, if feed control is managed solely based on the noisy measurements the time of feeding and number of feeding times deviate a lot from the ideal scenario (yellow, Figure 5D) leading to larger spikes in the resulting GLC profile (Figure 5C). Bad feeding decisions could adversely affect product quality and process performance. For instance, the two adjacent feed pulses stimulated by the measurement-based control between 200-250 h in Figure 5D could create a condition of overfeeding in the cell culture leading to increased lactate production and osmolality, detrimental effect on glycosylation and reduced productivity (Gilbert, Huang, & Ryll, 2014; Lim & Shin, 2013). In contrast, a situation of starvation could also occur leading to reduced cell growth and productivity, and alter glycosylation patterns (Fan et al., 2015).

3.4 Process Monitoring using Hybrid-EKF: Real Bioreactor Use Case

The applications shown so-far demonstrate the use of the Hybrid-EKF framework when all the states are measured in real-time. This is typically the situation when spectral measurements (Kozma et al., 2017) can be used to infer the real-time concentration of the metabolites through calibration models. Another interesting application are soft sensors, i.e., using these models to estimate an unmeasured, but

very relevant state such as product titer based on measurements of all the other states. Following this approach, the framework is applied next to the benchtop scale real bioreactor dataset (3.5 L) where the aim is to use Hybrid-EKF framework to estimate the titer with an accuracy comparable to the process error estimated based on replicate runs, which is about 40 mg/mL for titer. In a real application, the true values of the states are unknown since they are always observed with an error. The EKF for this case is thus tuned by correcting all states except titer to minimize the error between measured and estimated titer (since the aim is to build soft sensor for titer). What is worth noticing is that the use case is targeted to demonstrate a scenario, where the model is trained on an incomplete data set, which does not fully cover the possible process behavior and shall serve as basis to test for the extrapolation capability of such model. Similar to our earlier work (Narayanan, Sokolov, Morbidelli, et al., 2019), a model is trained on low titer experiments and applied to conditions leading to higher titers. The split of the data into calibration, tuning and test set can be seen in Figure 2A. First, it shall be evaluated how many experiments are required to optimize the EKF parameters to the new system. Figure 6A shows the evolution of EKF tuning function with the number of experiments from the new system. In order to make a fair comparison of the RMSEP, a fixed test set with 18 experiments is used. It can be observed that it is sufficient to have about 7 experiments to tune the EKF completely for the new system, after which the tuning objective remains almost constant. However, it is important to point out that, the framework can be used with sub-optimal EKF tuning and iteratively updated during the initial batches. Figure 6B through D shows the evolution of Titer, X_v and GLC for an example run from the test set as predicted by the Hybrid model alone (red

dash), the measurements acquired (yellow cross), the predicted state of Hybrid-EKF (blue) and the corrected states of Hybrid-EKF (green). It is interesting to note that the EKF chooses to trust the model over the measurements for GLC whereas for X_v, it is intermediate between the two information sources, in order to make an accurate estimation of Titer. It is reminded that all the other states are corrected using the measurements except for titer, while the filter parameters were tuned to make an accurate estimation of titer, which is unmeasured. The titer measurements are only used for evaluation of the RMSEs. Additionally, the comparison is also presented for the estimation of the titer by a classical historical PLS1 model for the titer. Essentially, a PLS1 model takes all the historical measurements until the current instance for all the states (except titer) and predicts titer (in other words, a PLS based soft sensor for titer). Similar to the hybrid, it is trained on low titer experiments and used to predict high titer conditions. It is evident that such industrial benchmark tool for titer estimation fails clearly producing an average RMSEP of 101.69 mg/L in comparison to the pure hybrid model with an average RMSEP of 58 mg/L on the entire test set. Further the Hybrid-EKF improves the state estimation, reducing the average RMSEP to 48 mg/L.

4 Discussion

The biopharmaceutical industry faces the need to meet increasing market demand, fulfill stringent quality regulations as well as digitalize and automate their procedures. Real-time process monitoring and control are key to tackle these goals. With the limitations in available generalized first principle models and diverse data

sets, hybrid modeling is emerging as a promising process modeling tool for cell culture processes.

In this work, we evaluate a generalized framework coupling two existing techniques, the hybrid models and Extended Kalman Filter (Hybrid-EKF) for application to mammalian cell culture through two process use cases. We show that appropriate selection of model and filter parameter are crucial for the successful implementation of the EKF. Correspondingly, we illustrate that for the presented use case using a hybrid model in such a framework improves the predictive accuracy by about 35% for the key variables, namely, X_v, Titer, GLC and LAC, when compared to using the industrial standard PLS algorithm. Additionally, it is also shown that the PLS models do not provide a strong model backbone to the EKF resulting in considerable dependence on the measurements. However, measurements are typically noisy and taking decisions based on them could be detrimental. As an example, we show that risk of wrong feed control decisions can be considerably reduced when using the predictive support of Hybrid-EKF algorithm. Further, we show the superior performance of the algorithm as a soft sensor for titer when compared to a classical PLS1 model on an industrial bioreactor dataset. The high prediction accuracy and the robust soft sensing capabilities make the presented framework an attractive support for precise online monitoring of process variables and quality attributes, and a basis for adaptive control and automatic decision making. Additionally, the robustness of the framework for process monitoring during process development is highlighted by training the model on low titer runs and performing state estimation for high titer runs. Given the extrapolative capabilities of the model, the framework can be also used to dynamically optimize titer or CQAs during the development

phase. On the other hand, during production, the predictive accuracy and extrapolative capabilities can be used to anticipate abnormal behavior and suggest strategies to reduce risks. Thus, the Hybrid-EKF framework has the potential to serve across different phases of the biopharmaceutical lifecycle.

To further exploit the advantage, case studies are planned to demonstrate the performance of the framework with Raman data and potentially also quantify the capability to estimate CQAs. Additionally, advanced filtering algorithms such as particle filter is considered to further improve the performance of such a framework.

Acknowledgement

We thank Moritz von Stosch for providing valuable comments to improve the scientific presentation of the work.

References

Berry, B., Moretto, J., Matthews, T., Smelko, J., & Wiltberger, K. (2015). Cross-scale predictive modeling of CHO cell culture growth and metabolites using Raman spectroscopy and multivariate analysis. *Biotechnology Progress*, *31*(2), 566–577. https://doi.org/10.1002/btpr.2035

Bhatia, H., Mehdizadeh, H., Drapeau, D., & Yoon, S. (2017). In-Line Monitoring of Amino Acids in Mammalian Cell Cultures using Raman Spectroscopy and Multivariate Chemometrics Models. *Engineering in Life Sciences*, 55–61. https://doi.org/10.1002/elsc.201700084

Catlin, T., Lorenz, J., Sternfels, B., & Willmott, P. (2017). *A roadmap for a digital transformation*. *McKinsey Quarterly* (Vol. March).

Craven, S., Shirsat, N., Whelan, J., & Glennon, B. (2013). Process model comparison and transferability across bioreactor scales and modes of operation for a mammalian cell bioprocess. *Biotechnology Progress*, *29*(1), 186–196. https://doi.org/10.1002/btpr.1664

Dewasme, L., Goffaux, G., Hantson, A. L., & Wouwer, A. Vande. (2013). Experimental validation of an Extended Kalman Filter estimating acetate

concentration in E. coli cultures. *Journal of Process Control*, *23*(2), 148–157. https://doi.org/10.1016/j.jprocont.2012.09.004

Dors, M., Simutis, R., & Lübbert, A. (1995). Advanced Supervision of Mammalian Cell Cultures Using Hybrid Process Models. *IFAC Proceedings Volumes*, *28*(3), 72–77. https://doi.org/10.1016/s1474-6670(17)45604-7

Fan, Y., Jimenez Del Val, I., Müller, C., Lund, A. M., Sen, J. W., Rasmussen, S. K., … Andersen, M. R. (2015). A multi-pronged investigation into the effect of glucose starvation and culture duration on fed-batch CHO cell culture. *Biotechnology and Bioengineering*, *112*(10), 2172–2184. https://doi.org/10.1002/bit.25620

Feidl, F., Garbellini, S., Luna, M. F., Vogg, S., Souquet, J., Broly, H., … Butté, A. (2019). Combining mechanistic modeling and raman spectroscopy for monitoring antibody chromatographic purification. *Processes*, *7*(10). https://doi.org/10.3390/pr7100683

Feyo de Azevedo, S., Dahm, B., & Oliveira, F. R. (1997). Hybrid modelling of biochemical processes: A comparison with the conventional approach. *Computers & Chemical Engineering*, *21*, S751–S756. https://doi.org/10.1016/S0098-1354(97)87593-X

Flickinger, M. C., Pohlscheidt, M., Charaniya, S., Bork, C., Jenzsch, M., Noetzel, T. L., & Luebbert, A. (2013). Bioprocess and Fermentation Monitoring. *Encyclopedia of Industrial Biotechnology*. https://doi.org/10.1002/9780470054581.eib606.pub2

Franceschini, G., & Macchietto, S. (2008). Model-based design of experiments for parameter precision: State of the art. *Chemical Engineering Science*, *63*(19), 4846–4872. https://doi.org/10.1016/j.ces.2007.11.034

Gilbert, A., Huang, Y., & Ryll, T. (2014). Identifying and eliminating cell culture process variability. *Pharmaceutical Bioprocessing*, *2*(6), 519–534. https://doi.org/10.4155/pbp.14.35

Gnoth, S., Jenzsch, M., Simutis, R., & Lübbert, A. (2008). Control of cultivation processes for recombinant protein production: A review. *Bioprocess and Biosystems Engineering*, *31*(1), 21–39. https://doi.org/10.1007/s00449-007-0163-7

Hastie, T., Tibshirani, R., & Friedman, J. (2017). The Elements of Statistical Learning The Elements of Statistical Learning. https://doi.org/10.1198/jasa.2004.s339

Jazwinski, A. H. (1970). *Stochastic Process and Filtering Theory, Academic Press*. *A subsidiary of Harcourt Brace Jovanovich Publishers*. A subsidiary of Harcourt Brace Jovanovich Publishers.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

Journal of Fluids Engineering, Transactions of the ASME, *82*(1), 35–45. https://doi.org/10.1115/1.3662552

Kozma, B., Hirsch, E., Gergely, S., Párta, L., Pataki, H., & Salgó, A. (2017). On-line prediction of the glucose concentration of CHO cell cultivations by NIR and Raman spectroscopy: Comparative scalability test with a shake flask model system. *Journal of Pharmaceutical and Biomedical Analysis*, *145*, 346–355. https://doi.org/10.1016/j.jpba.2017.06.070

Krämer, D., & King, R. (2017). A hybrid approach for bioprocess state estimation using NIR spectroscopy and a sigma-point Kalman filter. *Journal of Process Control*. https://doi.org/10.1016/j.jprocont.2017.11.008

Li, B., Ray, B. H., Leister, K. J., & Ryder, A. G. (2013). Performance monitoring of a mammalian cell based bioprocess using Raman spectroscopy. *Analytica Chimica Acta*, *796*, 84–91. https://doi.org/10.1016/j.aca.2013.07.058

Lim, H. C., & Shin, H. S. (2013). Phenomena That Favor Fed-Batch Operations. In *Fed-Batch Cultures* (pp. 52–61). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139018777.005

Lübbert, A., & Simutis, R. (1994). Using measurement data in bioprocess modelling and control. *Trends in Biotechnology*, *12*(8), 304–311. https://doi.org/10.1016/0167-7799(94)90047-7

Mandenius, C. F., & Gustavsson, R. (2015). Mini-review: Soft sensors as means for PAT in the manufacture of bio-therapeutics. *Journal of Chemical Technology and Biotechnology*, *90*(2), 215–227. https://doi.org/10.1002/jctb.4477

Markana, A., Padhiyar, N., & Moudgalya, K. (2018). Multi-criterion control of a bioprocess in fed-batch reactor using EKF based economic model predictive control. *Chemical Engineering Research and Design*, *136*, 282–294. https://doi.org/10.1016/j.cherd.2018.05.032

Mehdizadeh, H., Lauri, D., Karry, K. M., Moshgbar, M., Procopio-Melino, R., & Drapeau, D. (2015). Generic Raman-based calibration models enabling real-time monitoring of cell culture bioreactors. *Biotechnology Progress*, *31*(4), 1004–1013. https://doi.org/10.1002/btpr.2079

Mohd Ali, J., Ha Hoang, N., Hussain, M. A., & Dochain, D. (2015). *Review and classification of recent observers applied in chemical process systems*. *Computers and Chemical Engineering* (Vol. 76). Elsevier Ltd. https://doi.org/10.1016/j.compchemeng.2015.01.019

Narayanan, H., Luna, M. F., von Stosch, M., Cruz Bournazou, M. N., Polotti, G., Morbidelli, M., … Sokolov, M. (2020). Bioprocessing in the Digital Age: The Role of Process Models. *Biotechnology Journal*, *15*(1), 1–10. https://doi.org/10.1002/biot.201900172

Narayanan, H., Sokolov, M., Butté, A., & Morbidelli, M. (2019). Decision Tree – PLS (DT - PLS) algorithm for the development of process - specific local prediction models. *Biotechnology Progress*, (November 2018), e2818. https://doi.org/10.1002/btpr.2818

Narayanan, H., Sokolov, M., Morbidelli, M., & Butté, A. (2019). A new generation of predictive models–the added value of hybrid models for manufacturing processes of therapeutic proteins. *Biotechnology and Bioengineering*.

Ohadi, K., Legge, R. L., & Budman, H. M. (2015a). Development of a soft-sensor based on multi-wavelength fluorescence spectroscopy and a dynamic metabolic model for monitoring mammalian cell cultures. *Biotechnology and Bioengineering*, *112*(1), 197–208. https://doi.org/10.1002/bit.25339

Ohadi, K., Legge, R. L., & Budman, H. M. (2015b). Development of a soft-sensor based on multi-wavelength fluorescence spectroscopy and a dynamic metabolic model for monitoring mammalian cell cultures. *Biotechnology and Bioengineering*, *112*(1), 197–208. https://doi.org/10.1002/bit.25339

Rouiller, Y., Solacroup, T., Deparis, V., Barbafieri, M., Gleixner, R., Broly, H., & Eon-Duval, A. (2012). Application of Quality by Design to the characterization of the cell culture process of an Fc-Fusion protein. *European Journal of Pharmaceutics and Biopharmaceutics*, *81*(2), 426–437. https://doi.org/10.1016/j.ejpb.2012.02.018

Schiff, S. J. (2012). *Neural Control Engineering: The Emerging Intersection Between Control Theory and Neuroscience*. MIT Press. Retrieved from https://books.google.com/books?id=P9UvTQtnqKwC&pgis=1

Schneider, R., & Georgakis, C. (2013). How to NOT make the extended kalman filter fail. *Industrial and Engineering Chemistry Research*, *52*(9), 3354–3362. https://doi.org/10.1021/ie300415d

Simon, D. (2006). *Optimal State Estimation: Kalman, H∞, and Nonlinear Approaches*. *Optimal State Estimation: Kalman, H∞, and Nonlinear Approaches*. John Wiley & Sons. https://doi.org/10.1002/0470045345

Simutis, R., Havlik, I., & Lübbert, A. (1992). A fuzzy-supported Extended Kalman Filter: a new approach to state estimation and prediction exemplified by alcohol formation in beer brewing. *Journal of Biotechnology*, *24*(3), 211–234. https://doi.org/10.1016/0168-1656(92)90033-6

Simutis, R., Havlik, I., & Lübbert, A. (1993). Fuzzy-aided neural network for real-time state estimation and process prediction in the alcohol formation step of production-scale beer brewing. *Journal of Biotechnology*, *27*(2), 203–215. https://doi.org/10.1016/0168-1656(93)90109-Z

Simutis, Rimvydas, & Lübbert, A. (2017). Hybrid approach to state estimation for bioprocess control. *Bioengineering*, *4*(1), 21.

https://doi.org/10.3390/bioengineering4010021

Sokolov, M., Ritscher, J., MacKinnon, N., Souquet, J., Broly, H., Morbidelli, M., & Butté, A. (2017). Enhanced process understanding and multivariate prediction of the relationship between cell culture process and monoclonal antibody quality. *Biotechnology Progress*, *33*(5), 1368–1380. https://doi.org/10.1002/btpr.2502

Solle, D., Hitzmann, B., Herwig, C., Pereira Remelhe, M., Ulonska, S., Wuerth, L., … Steckenreiter, T. (2017). Between the Poles of Data-Driven and Mechanistic Modeling for Process Operation. *Chemie-Ingenieur-Technik*, *89*(5), 542–561. https://doi.org/10.1002/cite.201600175

Sommeregger, W., Sissolak, B., Kandra, K., von Stosch, M., Mayer, M., & Striedner, G. (2017). Quality by control: Towards model predictive control of mammalian cell culture bioprocesses. *Biotechnology Journal*, *12*(7), 1–7. https://doi.org/10.1002/biot.201600546

Svrcek, W. Y., Elliott, R. F., & Zajic, J. E. (1974). The extended Kalman filter applied to a continuous culture model. *Biotechnology and Bioengineering*, *16*(6), 827–846. https://doi.org/10.1002/bit.260160610

Teixeira, A. P., Oliveira, R., Alves, P. M., & Carrondo, M. J. T. (2009). Advances in on-line monitoring and control of mammalian cell cultures: Supporting the PAT initiative. *Biotechnology Advances*, *27*(6), 726–732. https://doi.org/10.1016/j.biotechadv.2009.05.003

Tie Jong, S. (1993). SIMPLS: an alternative approach squares regression to partial least. *Elsevier Science Publishers B.V*, *18*, 2–263. https://doi.org/10.1016/0169-7439(93)85002-X

von Stosch, M., Davy, S., Francois, K., Galvanauskas, V., Hamelink, J. M., Luebbert, A., … Glassey, J. (2014). Hybrid modeling for quality by design and PAT-benefits and challenges of applications in biopharmaceutical industry. *Biotechnology Journal*, *9*(6), 719–726. https://doi.org/10.1002/biot.201300385

Von Stosch, M., Oliveira, R., Peres, J., & Feyo De Azevedo, S. (2012). A general hybrid semi-parametric process control framework. *Journal of Process Control*, *22*(7), 1171–1181. https://doi.org/10.1016/j.jprocont.2012.05.004

Voss, J. P., Mittelheuser, N. E., Lemke, R., & Luttmann, R. (2017). Advanced monitoring and control of pharmaceutical production processes with Pichia pastoris by using Raman spectroscopy and multivariate calibration methods. *Engineering in Life Sciences*, *17*(12), 1281–1294. https://doi.org/10.1002/elsc.201600229

Wang, J., Zhao, L., & Yu, T. (2010). On-line estimation in fed-batch fermentation process using state space model and unscented kalman filter. *Chinese Journal*

of Chemical Engineering, *18*(2), 258–264. https://doi.org/10.1016/S1004-9541(08)60351-1

Xing, Z., Bishop, N., Leister, K., & Li, Z. J. (2010). Modeling kinetics of a large-scale fed-batch CHO cell culture by markov chain monte carlo method. *Biotechnology Progress*, *26*(1), 208–219. https://doi.org/10.1002/btpr.284

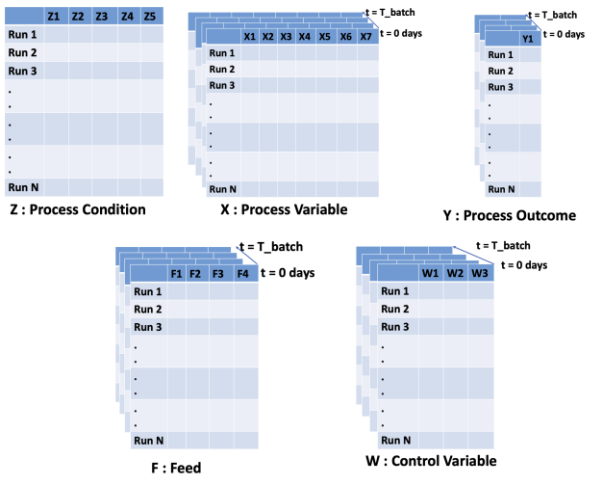
Zhang, H. (2009). Software sensors and their applications in bioprocess. *Studies in Computational Intelligence*, *218*, 25–56. https://doi.org/10.1007/978-3-642-01888-6_2

Zorzetto, L. F. M., & Wilson, J. A. (1996). Monitoring bioprocesses using hybrid models: Key features in Saccharomyces Cerevisiae production. Previous work on state estimation in Saccharomyces Cerevisiae production., *20*(96), 689–694.

Figures

Figure 1: (a) Data organization and nomenclature of different information matrices adapted from (Narayanan, Sokolov, Morbidelli, et al., 2019). (b) Compilation of the key features of the two analyzed datasets.

A



B

	INSILICO DATASET	REAL DATASET
Scale	Unit Volume	Bio Reactor (3.5 L)
Product	mAb	Fusion Fc protein
Medium	Chemically defined	Peptone - based
Mode	Fed - batch	Fed - batch
Culture duration (days)/ T_batch	14	10
Sampling	Every 2.4 h	Once per day
Process Condition (Z)	None	Amplification duration, Amplification VCD, Amplification cell age DO, pHset
Process variables (X)	VCD, GLC, LAC, GLN, NH4	VCD, GLC, LAC GLN, GLU, NH4, OSM
Feed Species (F)	GLC, GLN	GLC
Feed Start, End	Designed	Day 3 to Day 9
Controllable variable (W)	pH, T	pH, pCO_2, pO_2
Shifts	pH, T	None
Process outcome (Y)	Titer	Titer
States of System	X, Y	X,Y
# of runs	100	81
% missing X data	None	4

Figure 2: (a) Representation of dataset splitting into model training, EKF tuning and external test set (b) Schematic structure of the EKF implementation and potential applications in monitoring and control marked in different colors.

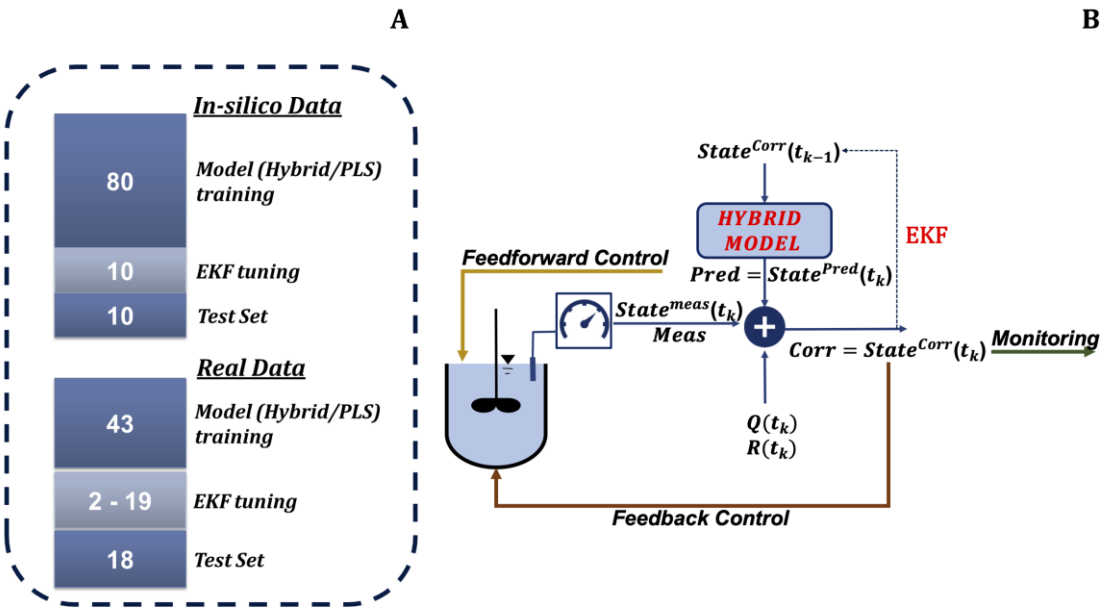


Figure 3: Time resolved absolute RMSEP calculated at discrete measurement point for the four key state variables namely, (a) X_v, (b) Titer, (c) GLC and (d) LAC compared for the measurement (blue), model (red), Hybrid-EKF module when k_Q is not optimized (yellow) and Hybrid-EKF module when k_Q is optimized (purple). RMSEP is computed with respect to true state values (simulated dataset).

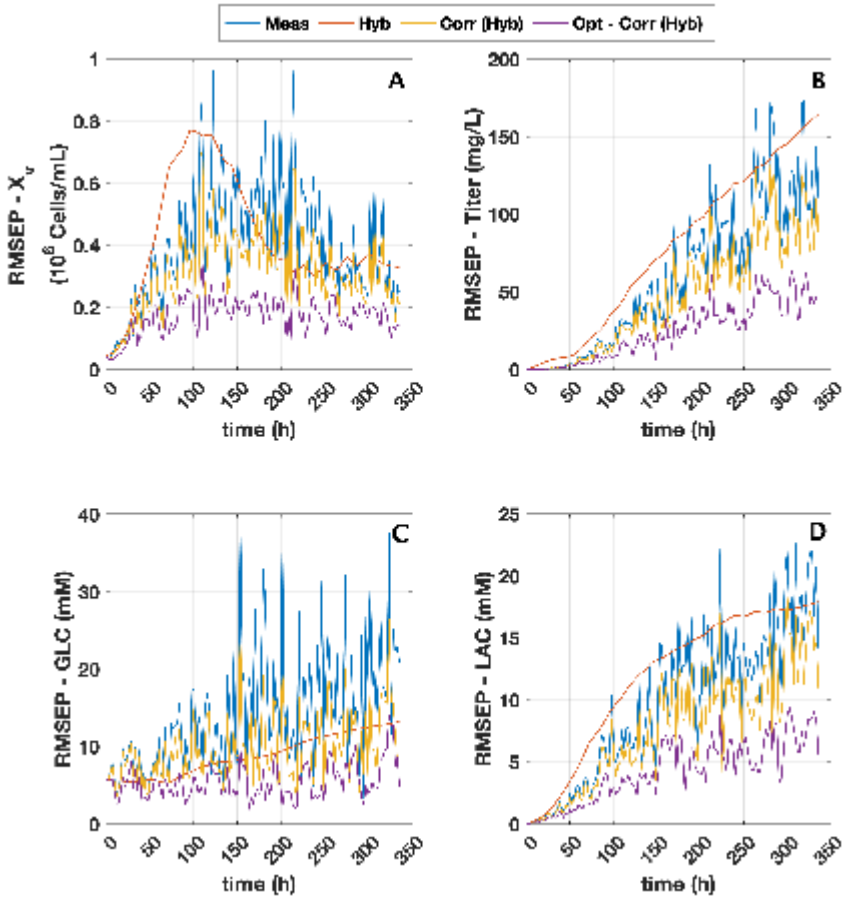
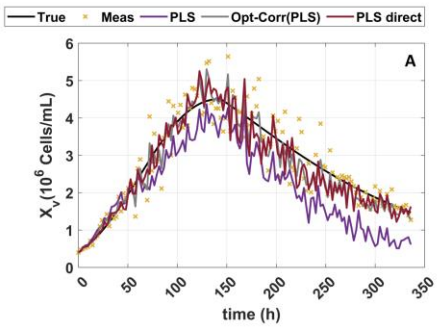
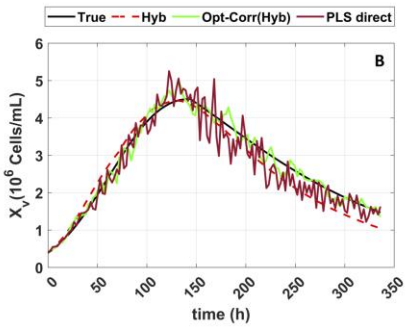
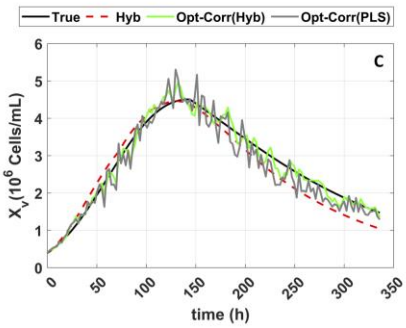


Figure 4: (a) Comparison of the true X_v (black) profile, noisy measurement (yellow cross), estimation of PLS-propagated model (purple, PLS), corrected state estimated by PLS-EKF module (grey, Opt-Corr(PLS)) and PLS – direct prediction (crimson, PLS direct) for an example run (b, c) Comparison of the true X_v profile, hybrid model prediction (red dash, Hyb), corrected states estimated by Hybrid-EKF (green, Opt-Corr(Hyb)) against PLS-direct and PLS-EKF module, respectively (d) Tabulates the absolute RMSEP (averaged over all times and test experiments) by the PLS-direct, PLS-EKF, Hybrid-EKF and Hybrid for X_v, Titer, GLC and LAC.







D

Variables	RMSEP			
	PLS-direct	PLS-EKF	Hybrid-EKF	Hybrid
X_v (10^6 Cells/mL)	0.27	0.29	0.18	0.47
Titer (mg/L)	50.50	45.60	30.87	94.13
GLC (mM)	8.83	12.78	5.9	9.27
LAC (mM)	6.84	6.34	4.95	13.18

Figure 5: (a) Comparison of the true GLC (black) profile, noisy measurement (yellow cross), estimation of hybrid model (red dash, Hyb), predicted state estimated by Hybrid-EKF module (blue, Opt-Pred(Hyb)) and corrected state estimated by Hybrid-EKF module (green, Opt-Corr(Hyb)) for an example run (b,c) GLC profile for a condition feeding experiment comparing the true feed strategy and feeding suggested by Hybrid-EKF and measurements respectively (d) Comparison of the actual feeding strategy with the feed triggered by Hybrid-EKF and pure noisy measurement.

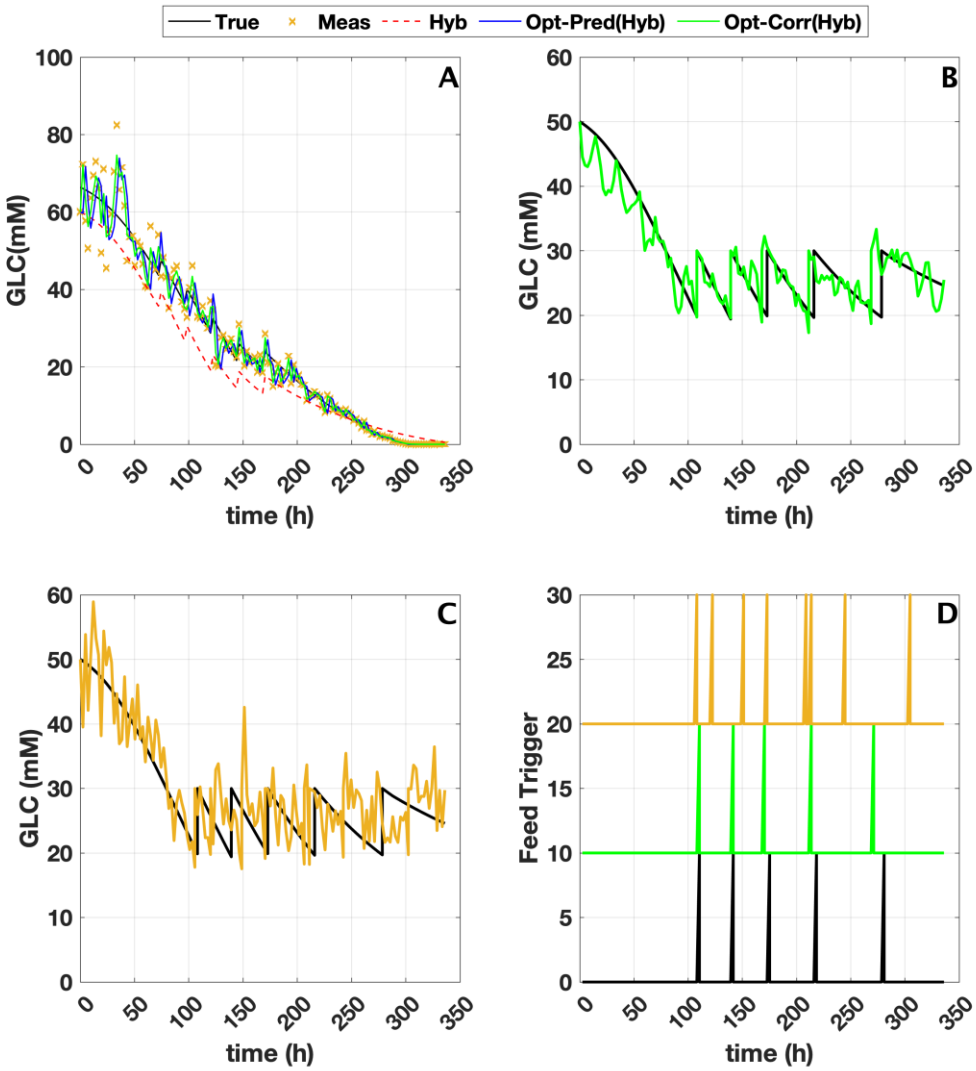


Figure 6: (a) Number of experiments required to tune the EKF of the Hybrid-EKF framework when transferring model across processes (b, c, d) Comparison of measurements (yellow cross), hybrid model prediction (red dash), predicted (blue) and corrected (green) state estimation by Hybrid-EKF for titer, X_v and GLC, respectively. Additionally, in (b) PLS1 based soft-sensor estimations for titer is also provided.