Received: 24 January 2019 | Revised: 13 May 2019 | Accepted: 18 June 2019

DOI: 10.1002/bit.27097

ARTICLE

Biotechnology and Bioengineering WILEY

A new generation of predictive models: The added value of hybrid models for manufacturing processes of therapeutic proteins

Harini Narayanan[1] | Michael Sokolov[1,2] | Massimo Morbidelli[1,2] | Alessandro Butté[1,2]

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

[2]DataHow AG, Zurich, Switzerland

Correspondence
Massimo Morbidelli, Department of Chemistry and Applied Biosciences, Institute of Chemical and Bioengineering, Vladimir-Prelog-Weg 1, 8093 Zürich, Switzerland.
Email: massimo.morbidelli@chem.ethz.ch

Abstract

Due to the lack of complete understanding of metabolic networks and reaction pathways, establishing a universal mechanistic model for mammalian cell culture processes remains a challenge. Contrarily, data-driven approaches for modeling these processes lack extrapolation capabilities. Hybrid modeling is a technique that exploits the synergy between the two modeling methods. Although mammalian cell cultures are among the most relevant processes in biotechnology and indeed looks ideal for hybrid modeling, their application has only been proposed but never developed in the literature. This study provides a quantitative assessment of the improvement brought by hybrid models with respect to the state-of-the-art statistical predictive models in the context of therapeutic protein production. This is illustrated using a dataset obtained from a 3.5 L fed-batch experiment. With the goal to robustly define the process design space, hybrid models reveal a superior capability to predict the time evolution of different process variables using only the initial and process conditions in comparison to the statistical models. Hybrid models not only feature more accurate prediction results but also demonstrate better robustness and extrapolation capabilities. For the future application, this study highlights the added value of hybrid modeling for model-based process optimization and design of experiments.

KEYWORDS

biopharmaceuticals, hybrid models, mammalian culture, predictive modeling, process optimization

1 | INTRODUCTION

With increasing market pressure to develop bioprocesses efficiently using minimal resources, biopharmaceutical industries have diverted their attention to model-based methods to aid understanding and decision-making at various stages of process development. Such activities are also supported by the health authorities through Process Analytical Tools (PAT) and Quality by Design (QbD; FDA, 2004; Mercier, Diepenbroek, Wijffels, & Streefland, 2014; Simon et al., 2015) initiatives motivating the use of quantitative tools for process understanding and quality control. Model-based methods have proven useful for bioprocess development in tasks such as the design of experiments, soft sensing, and spectral data processing by guiding experimental efforts in relevant regions, decreasing analytical costs of difficult-to-measure quantities, and extracting detailed information, respectively. In addition, model-based methods assist process design and optimization through simulated results and digital twins and support in scale-up, tech transfer, and improved process monitoring and control (Kroll, Hofer, Ulonska, Kager, & Herwig, 2017).

Models are mathematical approximations of the real-world phenomena that enhance understanding of the process, aids interpretation, and simplifies decision making. Depending on the amount of available *a priori* knowledge about the process and the data accessible for model development, a plethora of modeling techniques from purely statistical to purely mechanistic models can be formulated. The selection of the type of model depends on the goal of the modeling activity but is often highly biased by the expertise of the responsible team, available (commercial) software solutions and general resources such as time and labor. However, the essential qualities of models are the ability to describe the relevant reality with sufficient accuracy, the capability for model transfer to different situations (model robustness) and the simplicity of its development and interpretation (Solle et al., 2017).

Black-box (or data-driven) models rely completely on data and captures relevant process behavior solely based on statistical correlations between the input and output variables. Depending on the nature of the algorithm, such models can be generated quite efficiently and in sufficiently simple cases with limited expertise. A substantial amount of good quality data is needed to train a reliable model, which however is usually not valid in regions that were not explored in the underlying experiments. These models can be used to evaluate the variable importance and correlations, recognize the sources of process noise and abnormality and elucidate nonintuitive interrelationships so to hypothesize on new insights. Nonetheless, such models do not generate new process knowledge.

On the contrary, there are the first principle models (FPMs), which are derived from physical, chemical, and biological principles. In process engineering, such models are constituted by mass and energy balances, thermodynamics, transport phenomena, and reaction kinetic schemes expressed as a mixed system of differential (ordinary or partial) algebraic equations. Compared to those in the data-driven model, the parameters in FPMs bear a distinct physical meaning and therefore can often be estimated *a priori*, independent of the specific process under consideration. As long as they capture the relevant phenomena, FPMs are highly reliable and show good extrapolation capabilities. However, not only is such model generation time-consuming but it is also not possible to create FPMs when the underlying phenomena are not fully understood or the measurements of variables required for parameter estimation are not available.

An attractive synergy between the two approaches can be established by the so-called hybrid models. The concept of this modeling technique relies on establishing a mechanistic model framework whereas using data-driven approaches to estimate the unknown parts of the equations and flexibly adapt it to different scenarios. The embedded mechanistic structure improves model robustness and extrapolation while reducing overfitting and the amount of data required. The data-driven part simplifies the management of system complexity and the estimation of model parameters and sensitivity. A detailed description of application, definition, advantages, and disadvantages of all the three modeling techniques is reported in a recent general review by (Solle et al., 2017) and from the perspective of QbD and PAT by (Simon et al., 2015; Teixeira, Oliveira, Alves, & Carrondo, 2009; Von Stosch et al., 2014).

There can be different possible architectures for developing hybrid models as described in detail in (Solle et al., 2017; Thompson & Kramer, 1994). These are referred to as serial when using the Black-box models to estimate unknown terms in the mechanistic equations, parallel when using the Black-box model to reduce the errors made by the mechanistic model or a hybrid approach when the mechanistic model is used to generate data for training Black-box models. The earliest literature in the direction of hybrid models dates back to 1992 in the field of biotechnology and to 1999 in Chemical engineering. In bioprocess engineering this included general bioreactor modeling (Psichogios, 1992) and production processes for penicillin (Can, Braake, Hellinga, Luyben, & Heijnen, 1997; Montague et al., 2010; Thompson & Kramer, 1994), baker's yeast (Feyo de Azevedo, Dahm, & Oliveira, 1997; Oliveira, 2004; Schubert, Simutis, Dors, Havlik, & Lubbert, 1994a, 1994b), and beer (Zorzetto, Filho, & Wolf-Maciel, 2000). Hybrid models have also been reported for several applications in chemical engineering (Georgieva, Feyo de Azevedo, Gonçalves, & Ho, 2003; Hu, Mao, He, & Yang, 2011; Nagrath, Messac, Bequette, & Cramer, 2004; Tian, Zhang, & Morris, 2001; Zander & Dittmeyer, 1999; Zhang, Mao, Jia, & He, 2015). Differences among these applications relate not only to the model architecture but also to the algorithm used in the data-driven parts. Although in most cases artificial neural networks (ANNs) were used some applications of support vector regression (Hu et al., 2011; Yang, Martin, & Morris, 2011), nonlinear partial least square (PLS; Von Stosch, Oliveira, Peres, & Feyo de Azevedo, 2011), and other Black-box models (Tian et al., 2001) can also be found.

In recent years, hybrid models have been used successfully for model predictive control (Sommeregger et al., 2017; Von Stosch, Oliveira, Peres, & Feyo de Azevedo, 2012), process monitoring and forecasting (Von Stosch, Hamelink, & Oliveira, 2016; Zorzetto & Wilson, 2003), iterative process optimization (Teixeira et al., 2005; Teixeira, Clemente, Cunha, Carrondo, & Oliveira, 2006; Teixeira, Alves, Alves, Carrondo, & Oliveira, 2007) and also in the downstream chromatographic processes (Creasy, Barker, Yao, & Carta, 2015) in the context of therapeutic protein manufacturing. The use of these techniques in the mammalian cell culture bioreactors has been conceptualized in a recent publication (Sommeregger et al., 2017) but is demonstrated quantitatively for the first time in this study. In particular, we develop a hybrid process model, based on ANNs and mass balance equations, to predict the time evolution of the most relevant state variables in a fed-batch mammalian cell culture bioreactor for the production of monoclonal antibody. The performance of the obtained hybrid model is compared to the state-of-the-art statistical models in terms of model accuracy, interpolation, and extrapolation capabilities and potential application in process optimization and design of experiments have been demonstrated.

2 | MATERIALS AND METHODS

2.1 | Dataset

The hybrid model developed in this study was tested on a cell culture process dataset originally published by (Rouiller et al., 2012). 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 and two process conditions being pH and dissolved oxygen set points. An extensive study varying the amplification and process condition in wide ranges were performed. The amplification N-1 process cell density was varied between 5.51×10^6 and 7.17×10^6 cells/ml whereas the cell age was varied between 23 and 35 days. The amplification duration was either 4 or 5 days. Among the process conditions, the pH set point was varied from 6.7 to 7.2, whereas the DO set points were varied from 10% to 70%. A detailed analysis of the dataset can be found in Sokolov et al. (2015). For the modeling purpose, this information will be denoted as matrix Z representing all designed conditions. In general, Z is a two-dimensional matrix with rows and columns representing runs and manipulated variables, respectively. The dynamically changing, non-controlled process variables namely viable cell density (X_v), concentration of glucose (GLC), lactate (LAC), glutamine (GLN), glutamate (GLU) and ammonia (NH_4), and osmolality (Osm) were measured every single day from start until the end of the run (10 days). This provides us with a three-dimensional matrix X, which features an additional time dimension in comparison to the Z matrix. Since the metabolite concentrations are measured daily just before the bolus feed addition, the mass of different metabolites added to the fed-batch during the culture time is organized into a different matrix referred as F which is obviously a three-dimensional matrix with same dimensions as X. During these cell culture experiments, bolus glucose feed was supplied daily to the reactors starting from culture Day 3 onwards. An additional process information matrix, referred to as W, is built with all variables which are controlled to remain close to a certain set point. The W matrix thus includes pH, partial pressures of carbon dioxide (pCO_2) and of oxygen (pO_2) which were measured daily. Finally, product titer was measured on alternate days from Day 0 until the end of the run (i.e., Days 2, 4, 6, 8, and 10). Due to the extensive testing of different conditions, the final titer ranges from 200 to 700 mg/L and the final specific productivity varies between 1.5 and 14 pg cells/day. A logistic interpolation (Goudar, Joeris, Konstantinov, & Piret, 2005) for each run was performed to estimate the value of titer on days when it was not quantified that is on Days 1, 3, 5, 7, and 9. This information will be represented as matrix Y. Both W and Y have a three-dimensional structure similar to the X matrix.

The organization and dimensions of the different data matrix are illustrated schematically in Figure 1a, while the detailed description of the process set up, operation mode, and analytics procedure can be found in the original literature. The organization of the dataset is facilitated by the appearance of the different information in the hybrid model. Matrix Z is constant in time while X and W are dynamic information that is used in the Black-box part of the hybrid model. However, due to the presence of a mechanistic framework for computing X matrix terms, model concentrations are used in the Black-box model, whereas experimental measurements are used directly for the W variables. The relevant mass balances are integrated for each culture day and eventually, the feed addition is simulated to account for the new starting point of the next daily integration (discussed in detail in Section 2.2.1). Finally, Y is not used in the model computations since titer is actually computed by the model, which therefore acts as a soft sensor with respect to titer. The dataset contains about 4% fairly randomly distributed missing data which was imputed using the Trimmed Score Regression algorithm (Folch-Fortuny, Villaverde, Ferrer, & Banga, 2015).

2.2 | Methodology

The dataset is first divided into calibration and test set in the ratio of 80–20% before training the model. The different models are then trained on the calibration set and used to predict the test set. During calibration, a cross-validation strategy is used to tune the hyperparameters of the model. The performance of the different models is compared based on the root mean squared error made in prediction (RMSEP). All computations are performed using MATLAB R2017a.

2.2.1 | Hybrid model

A serial architecture hybrid modeling approach (Thompson & Kramer, 1994) as illustrated in the schematic flow diagram in Figure 1b is adopted in the current work. The system of equation representing the cell culture is established based on mass balances, reported as follows:

$$\frac{dC_i}{dt} = \mu_i(t)C_1(t) \text{ with } T^{+} \le t < T + 1, \tag{1}$$

such that T varies from culture Days 0 to 9 with T^{+} indicating the time point just after bolus feed addition and where C_i and μ_i are the concentrations and specific rates of the ith species, respectively, and i represents X_v, GLC, LAC, GLN, GLU, NH_4, Osm, and titer. Therefore, there are eight modeling targets with C_1 representing the $X_{v.}$ Due to the lack of complete understanding of cell metabolism, the specific rates are often unknown up front. This lack of knowledge is compensated with the Black-box model as shown in the schematics in Figure 1b, which estimates specific rates based on the information from the culture experiments.

$$\mu(t) = f(X_v, \text{GLC}(t), \ \text{LAC}(t), \ \text{GLN}(t), \ \text{GLU}(t), \ \text{NH}_4(t), \\ \text{OSM}(t), \ Z, \ W(t)). \tag{2}$$

Any regression tool of the generic form stated in equation (2) can be used to map the process factors and process variable measurements to the μs. In particular, a feed-forward single hidden layer ANN is used in this study which automatically takes into account the different transformation of the inputs due to its nonlinear structure. The detailed explanation of the ANN formulation and activation functions can be found in the literature by (Von Stosch et al., 2016). Essentially, integration and optimization are performed simultaneously to optimize the neural network weights such that the difference between measured and model predicted concentration

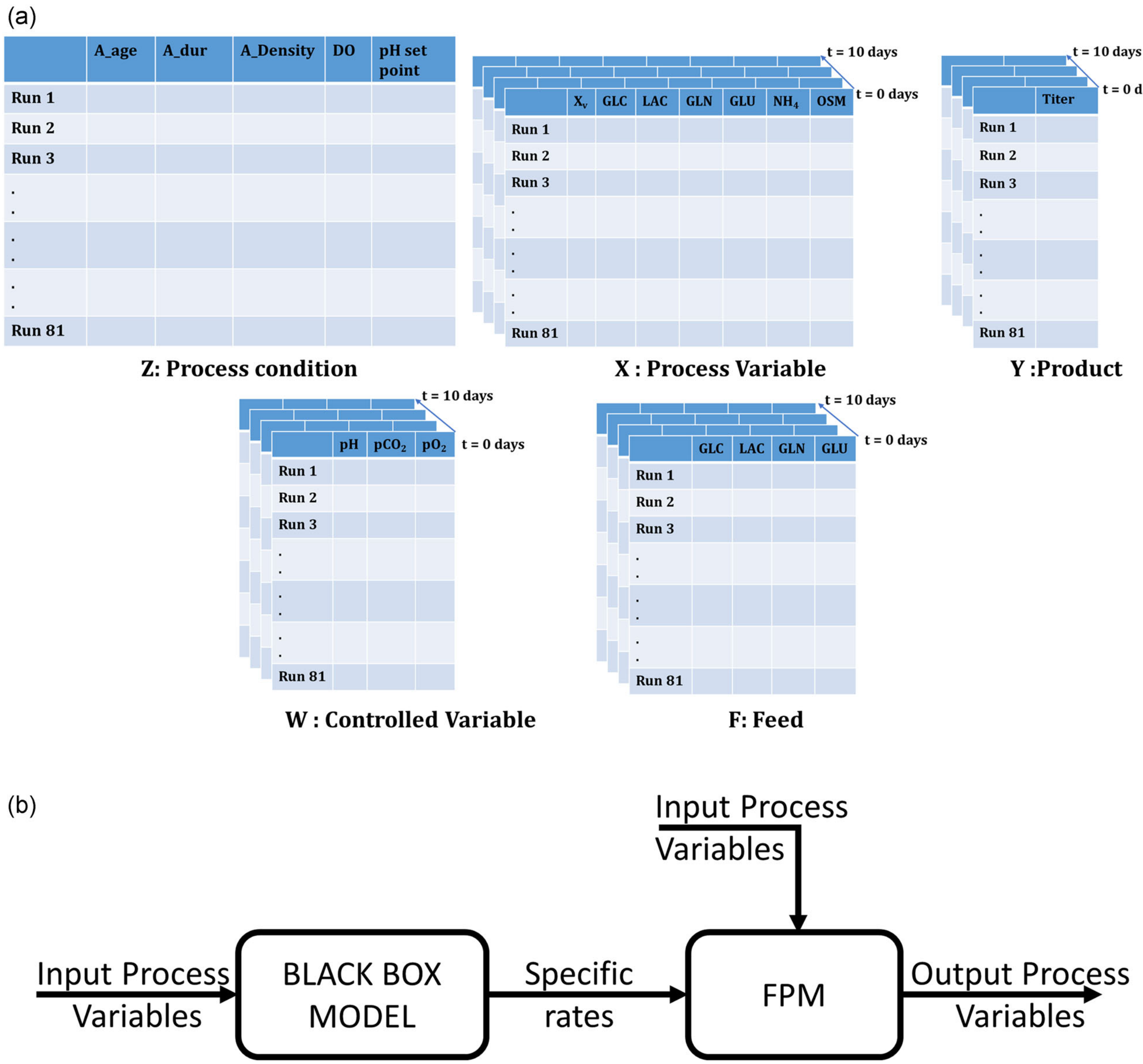


FIGURE 1 (a) Data organization and nomenclature of different information matrices. (b) Schematic structure of the serial hybrid model architecture. FPM, first principle models [Color figure can be viewed at wileyonlinelibrary.com]

values is minimized. Integration is performed for the time span of a day and eventually, the feed addition is simulated as follows:

$$C_{i,init}(T) = C_i(T) + \frac{Feed_i(T)}{V_R}, \tag{3}$$

where C_i (T) represents the concentration value before the feed addition, which also corresponds to the measured value, while $C_{i,init}$ (T) represents the initial condition for integrating Equation (1) from T to $T + 1$, $Feed_i$ (T) the mass of species i, that is fed, and V_R the reactor volume. A two-norm regularized objective function (Yang et al., 2011) is used to avoid overfitting of the weights in the neural network and five-fold cross-validation (Hastie, Tibsharani, & Friedman, 2009) is used to determine the optimal number of nodes and the regularization parameter. *ode15s ( )* is used for integration and *fminunc ( )* is used for optimization purposes.

2.2.2 | Statistical reference models

The performance of the hybrid model was compared with two statistical models, which were built based on the same initial information, that is information matrices X $(t = 0)$, W $(t = 0)$, Z, and F as used in the hybrid model for predicting the dynamic process behavior. The first model hereon referred to as BB–PLS2 is a black-box–PLS2 model predicting simultaneously all the eight target variables at model-specific time points using only the Z and initial conditions, that is, X $(t = 0)$ and W $(t = 0)$. As a result, there are 10 PLS2 models each calibrated on a different culture day. The second one, Stepwise–PLS2 is a single PLS2 model where X and W at a certain time T are used to predict the immediate daily change in the target variable values, $\Delta Y(T)$. In other words, the mapping can be represented as follows:

$$[X(T), W(T)] \rightarrow \Delta Y(T) = Y(T+1) - Y(T) \tag{4}$$

such that T varies from Days 0 to 9. Thus, the input and output to the Stepwise–PLS2 model essentially contain 81 × 9 (=729) rows, that is, the number of runs (81 runs) multiplied by the number of time points corresponding to each run (9 days). However, it is worth noting that during the prediction phase only input at $T = 0$ is given to the Stepwise–PLS2 model. Thereby, the value at time $T = 1$ is predicted by the model as shown in equation (5) below, which then serves as an input to the model for predicting the next time point $T = 2$ and so forth.

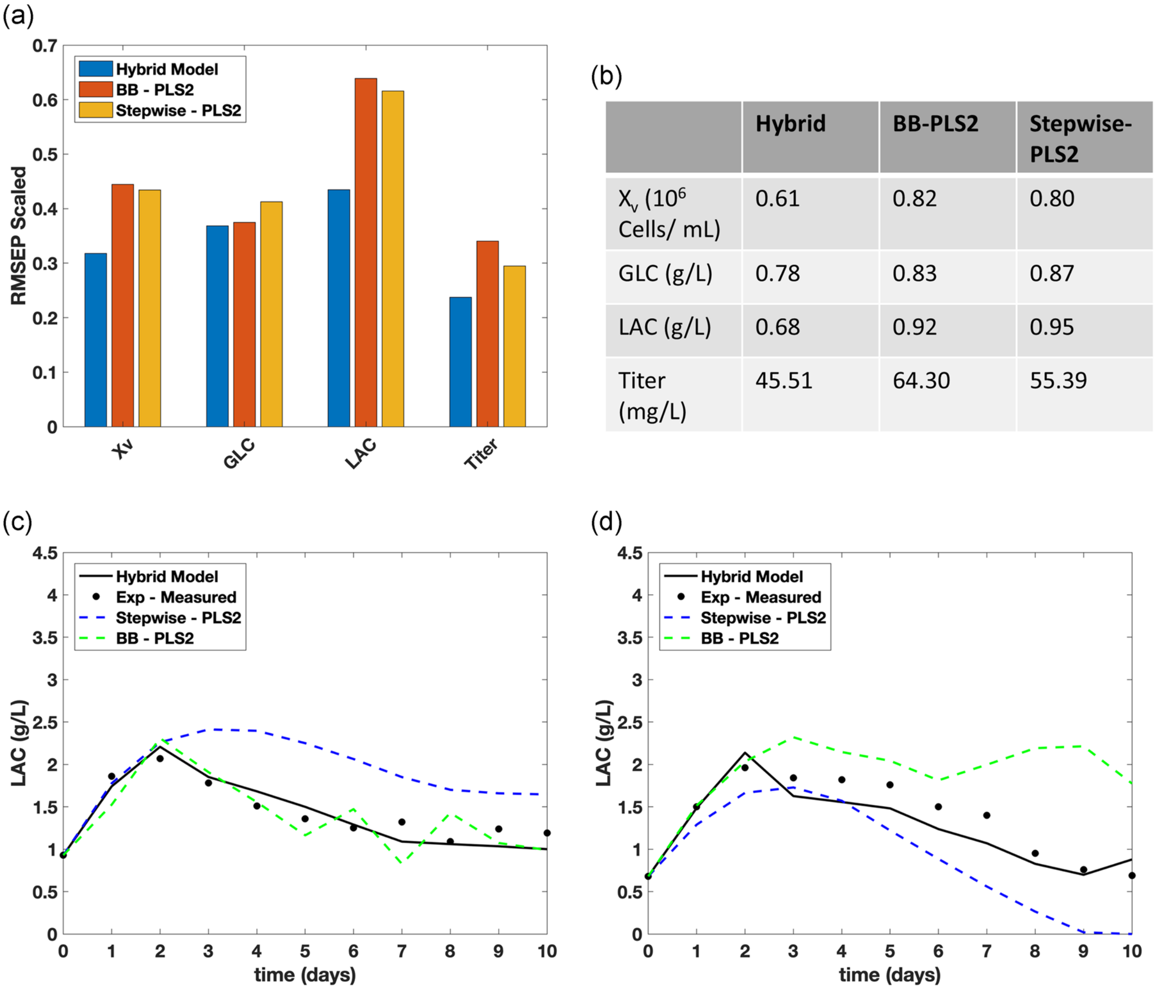
$$Y(T-1) = Y(T) + \Delta Y(T), \tag{5}$$

where $\Delta Y(T)$ is predicted by the Stepwise–PLS2 model. All the PLS models were developed using the in-built MATLAB function *plsregress ( )* which is based on the SIMPLS algorithm (Tie Jong, 1993). In all cases, a five-fold cross-validation strategy was used to select the optimal number of latent variables.

3 | RESULTS

3.1 | Model accuracy

The accuracy of hybrid model was compared to that of the statistical models in terms of the scaled RMSEP of the four key variables, namely, X_v, GLC, LAC, and titer, as shown in Figure 2a. The errors were scaled to the respective standard deviation of each variable computed across all the calibration runs and all time points. It is seen that the hybrid approach exhibits the lowest scaled RMSEP for X_v, LAC and titer among all the models compared. It is worth noting that as all the metabolites are processed by living cells, their estimates strongly depend on the one of X_v. Thus, it is crucial to predict X_v as accurately as possible. In Figure 2a it is seen that the hybrid model predicts X_v with a scaled error of 0.3 as compared to the statistical models which make an error of about 0.45. In terms of absolute errors, this corresponds to 0.61 and 0.80 × 10^6 cells/ml, respectively (Figure 2b). The error made by the hybrid model is close to the process analytics which is around 0.5 × 10^6 cells/ml. In the case of



	Hybrid	BB-PLS2	Stepwise-PLS2
X_v (10^6 Cells/ mL)	0.61	0.82	0.80
GLC (g/L)	0.78	0.83	0.87
LAC (g/L)	0.68	0.92	0.95
Titer (mg/L)	45.51	64.30	55.39

FIGURE 2 Comparison of the accuracy of the three models in terms of (a) root mean squared error made in prediction (RMSEP) scaled with respect to the standard deviation; (b) absolute RMSEP for viable cell density (X_v), glucose (GLC) and lactate (LAC) concentration, and titer. (c,d) Comparison of the lactate concentration predicted by the three models as a function of time for two exemplary runs with experimental data [Color figure can be viewed at wileyonlinelibrary.com]

lactate concentration, an immense improvement in prediction is achieved when using hybrid models with the scaled errors of about 0.45, whereas the BB–PLS2 and Stepwise–PLS2 models make a significantly higher scaled error of 0.65 and 0.7, respectively. An interesting insight in the different behavior of these models can be obtained by comparing the lactate concentration values predicted by the different models as a function of time shown in Figure 2c,d together with the experimental measurements for two exemplary experimental runs. Although both runs are predicted well by the hybrid model, Stepwise–PLS2 is unable to predict the actual profile from Day 2 onwards. On the contrary, the BB–PLS2 predicts one of the two exemplary runs well but the profile is not smooth in either case. This suggests that the BB–PLS2 model for lactate prediction is not very general and might have overfitted a certain type of runs.

Finally, for the target protein prediction, it can be observed from Figure 2a that the hybrid model predicts the titer with a scaled RMSEP of 0.2. As shown in Figure 2b, this corresponds to an absolute RMSEP of approximately 46 mg/L which is very close to the expected analytical error of about 30 mg/L. On the contrary, BB–PLS2 and Stepwise–PLS2 model make an absolute RMSEP of 65 and 55 mg/L for titer prediction. Specifically, as shown later in Figure 3d, the absolute RMSEP made by Hybrid model for the prediction of final day titer is around 80 mg/L which is much lower than those made by Stepwise–PLS2 and BB–PLS2 models that make an error of 100 and 120 mg/L, respectively. Such improved model accuracy in forecasting is crucial to provide reliable decision support to compare different experimental scenario and guide the design of the subsequent runs as well as for process optimization and control application. The higher precision of such models facilitates its use as a digital twin to simulate experiments with several operating conditions in silico and identify the optimal design space, whereas significantly reducing the experimental and analytical effort.

3.2 | Model robustness

The overall performance of Stepwise–PLS2 and BB–PLS2 is comparable as observed from the averaged RMSEP across all days shown in Figure 2 a. However, the time-resolved scaled RMSEP for the four key variables in Figure 3 suggest that the BB–PLS2 models make a

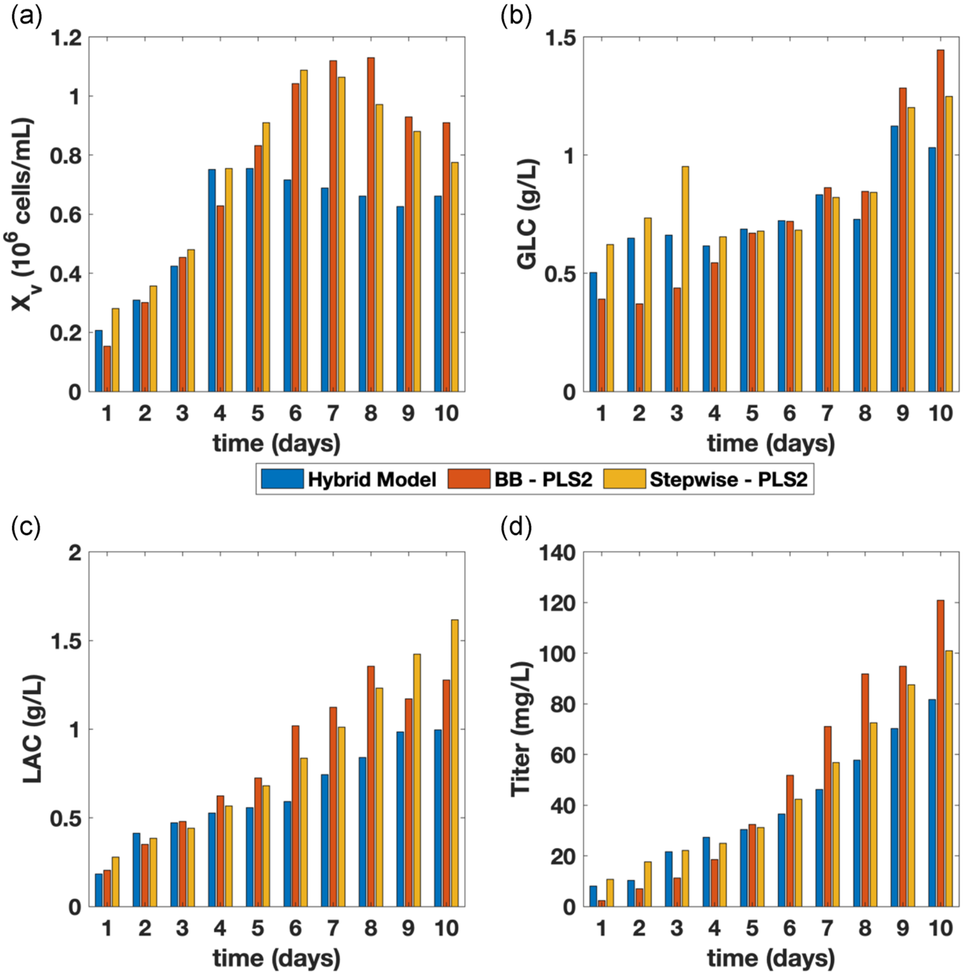


FIGURE 3 Time-resolved absolute root mean squared error made in prediction (RMSEP) made by the different models in the prediction of the four key variables, namely, (a) viable cell density (X_v, 10^6 cells/ml), (b) glucose (GLC, g/L) and (c) lactate (LAC, g/L) concentration, and (d) titer (mg/L). The *X*-axis distinguished the cell culture day that the model predicts, whereas the *Y*-axis shows the RMSEP made by the model (averaged over all test runs) for a certain variable at that time point [Color figure can be viewed at wileyonlinelibrary.com]

better prediction on early days as compared to both hybrid and Stepwise–PLS2 models, but for later days the errors increase significantly and exceed the RMSEP of the other two models. This can be understood based on the different nature of the models under consideration. The BB–PLS2 models estimate the concentration at different times during the fed-batch using only the initial conditions and the process operating conditions. As a result, these models can well estimate the state variables for a few days from the start of the culture but fails to extrapolate its behavior for longer times as they have no information about the overall dynamic evolution of the system. On the contrary, in both the Stepwise–PLS2 and hybrid model, a single model estimates the dynamic evolution of all the variables simultaneously. Thus, the scaled RMSEPs are not specifically small at the beginning of the culture but are rather evenly distributed as both models capture the time specific effects in a generalized model framework. In addition, as the concentrations at a certain time are predicted based on concentration values estimated at previous times, the corresponding error accumulates in time as shown in Figure 3 for titer and LAC concentration.

3.3 | Interpolation capability and flexibility

In contrast to the other variables, as discussed in the context of Figure 2a, overall model accuracy in predicting the glucose concentration is similar for all considered models. Nevertheless, the way different models treat the glucose dynamics during the process is remarkably different as shown in Figure 4a. This provides an interesting opportunity to appreciate the different potential of these models. On the one hand, the statistical model does not possess the capability of accessing, learning, and hence predicting what occurs in the system in between the measurement points. On the other hand, the solid mechanistic framework in the hybrid model allows it to predict the near-reality state of the system between the measurement points. The FPM is in fact based on mass balance equations, and therefore the feed conditions are accounted for in the hybrid model. Though they cannot be captured by the statistical model, given the measurements is usually taken before the feed additions. Accordingly, the hybrid model correctly predicts a continuous decrease in glucose concentration during each day, after the addition of glucose (from Day 3 onwards), whereas the statistical models feature an increase in glucose concentration. This highlights one of the key capabilities of hybrid models, that is to correctly follow the physics of the process and consequently exhibit better interpolation and extrapolation capabilities compared to the statistical models. This has important consequences when it comes to the use of such models for process simulation. For example, Figure 4b compares the same experimental run demonstrated in Figure 4a with a simulated run having a different feeding strategy. The original experiment (for which the experimental observation is plotted in Figure 4a) has a feeding scheme of 9 g daily glucose addition from culture Day 3 onwards, whereas the simulated run is presented for a 6 g glucose addition instead. For such subtle changes in the feeding scheme such as the concentration of feed added, the statistical models are incapable of accounting for the changes whereas only the hybrid

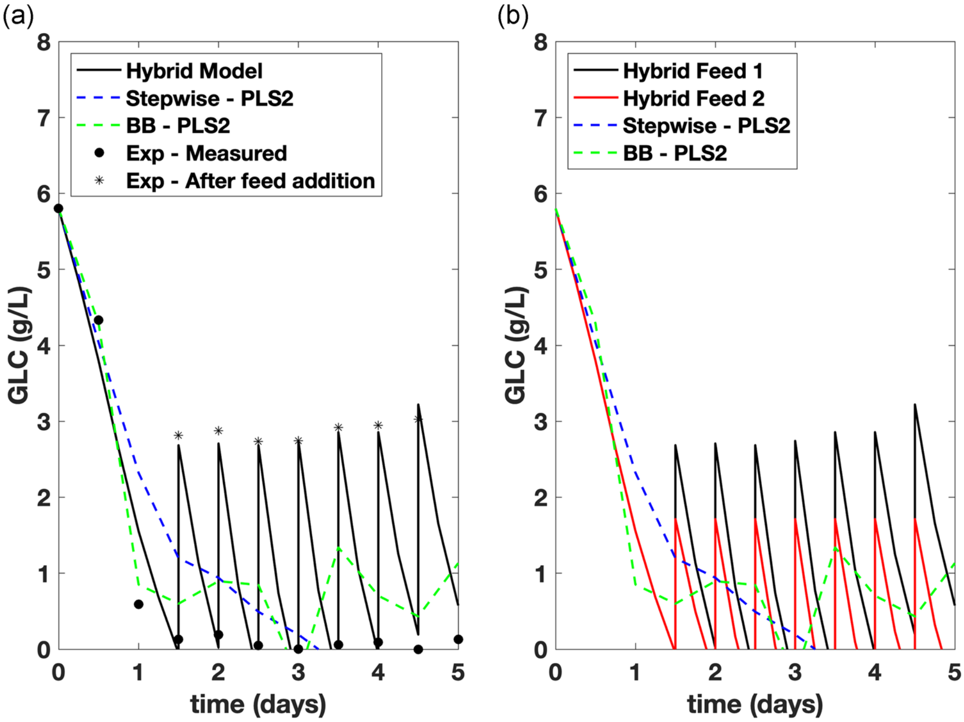


FIGURE 4 (a) Comparison of the glucose concentration as a function of time for an exemplary run predicted by the different models with experimental data; (b) glucose concentration as a function of time predicted by the different models for two different feeding strategies, one observed experimentally, same as (a), other simulated with a hypothetical feeding addition. BB–PLS2, black-box–partial least square 2 [Color figure can be viewed at wileyonlinelibrary.com]

models can distinguish the different feeding strategies (also in a physically relevant manner). Another advantage is that unlike the statistical models, which need a stringent time alignment of all measurements across the different runs, the hybrid model, due to the FPM, can provide estimates at any time point during the process dynamics. This allows for modeling bioreactor runs of different lengths and different sampling frequency using the same model.

3.4 | Process extrapolation and optimization

One of the key goals in biopharmaceutical process development is to identify process conditions and feeding strategies that can maximize productivity. This requires exploring a large number of possible operating conditions and their combination, which, if done experimentally, would involve high costs and long times. Just like in any other industrial sector, it is much more convenient to use in-silico experiments based on suitable mathematical models, to properly reduce the optimal design parameter space. This requires to ensure model reliability, particularly in predicting process performances not previously explored experimentally. To quantitatively illustrate this point, the three models were trained on experiments with a low final titer, that is, less than 580 mg/L and tested on the experiments with a titer greater than 580 mg/L. The cutoff was chosen such that at least 50% of the dataset (41 runs) could be used for training the models. In Figure 5a, a comparison of different models in predicting the experiments leading to high titer is shown. It is seen that the hybrid model exhibits the smallest error in predicting the four key variables, namely, X_v, GLC, LAC, and titer. In comparison to Figure 2a, the scaled RMSEPs made by the hybrid models have increased by 5–10%. For instance, for X_v, the errors increased from 0.3 to 0.35, whereas for titer from 0.2 to 0.3. However, the BB–PLS2 and Stepwise–PLS2 models make an error of 45–50%, respectively in titer prediction and as high as 65% and 80% in lactate prediction, respectively. This confirms that these models are much less reliable for performing in-silico experiments to identify appropriate conditions leading to high titer.

The conclusion is supported by the results in Figure 5b representing the X_v as a function of time for an exemplary test run predicted by the hybrid, BB–PLS2 and Stepwise–PLS2 model. As mentioned previously, the estimate of all the other variables strongly depends on the estimate of X_v. It can be observed that the BB–PLS2 and Stepwise–PLS2 that are trained on the low titer runs are incapable of accurately predicting X_v of the high titer runs. On the contrary, the hybrid model is capable of extrapolating beyond the trained region reasonably well as seen from its capability to predict the X_v profiles for high titer runs.

4 | DISCUSSION

The biopharmaceutical industry is currently facing two major issues. First, there is the innovation pressure driven by the patient needs for less expensive protein-based therapies. Though the other relates to digitalization pressure driven by the need to benefit from the immense size of generated data pools and the increasingly diversifying choice of data management and analytics solutions. A success in efficient data management and analytics is expected to have a tremendous effect on improving process development strategies and

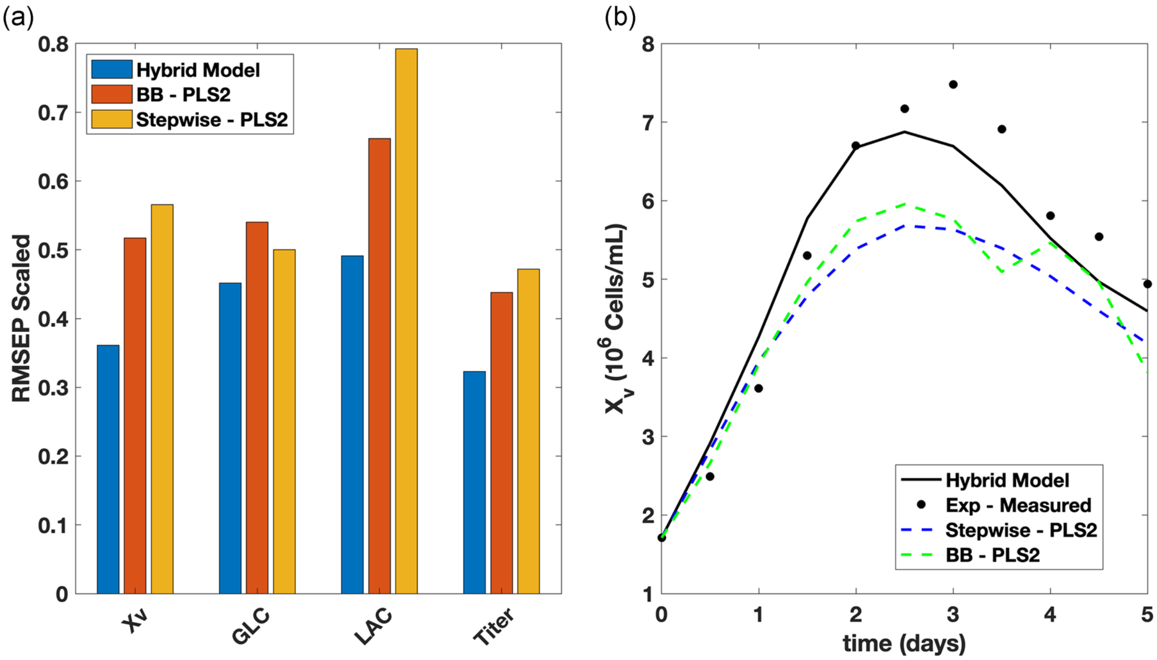


FIGURE 5 (a) Comparison of scaled root mean squared error made in prediction (RMSEP) exhibited by the different models trained on low titer runs when predicting the high titer runs; (b) comparison of viable cell density (X_v) for an exemplary run predicted as a function of time by the Hybrid, black-box–partial least square 2 (BB–PLS2), and Stepwise–PLS2 models with experimental data [Color figure can be viewed at wileyonlinelibrary.com]

reducing manufacturing costs, thus making these therapies more available to the society

On the one hand, the lack of complete understanding of the dynamics of cell cultures that limits the progress of FPMs. On the other hand, statistical models, solely based on data, perform well only in the regions of the operating conditions that have been investigated experimentally. However, it is important to have process models that are accurate in prediction and capable of extrapolating beyond trained conditions. In this study, we have demonstrated the potential of the hybrid modeling approach in this direction, with reference to the relevant case of the fed-batch production of a commercial monoclonal antibody.

In particular, we have shown that hybrid models provide more accurate predictions of several key variables including metabolite concentrations, viable cell density, and titer with respect to purely statistical approaches. It is remarkable that this result is obtained using a mechanistic framework based on very simple mass balance equations. Nevertheless, such a modest modeling effort is sufficient to generate a hybrid model capable of predicting the complete evolution of the process variables and titer solely based on the initial conditions and the process operating conditions. Hybrid modeling also shows good interpolation and extrapolation capabilities in comparison to statistical models, which suffer from time alignment problems and face intrinsic limitations to predict outside their training operation region. This extrapolation capability is not only central for the model-based process optimization but also for process scale-up, technology transfer and decision-making in process development.

An important limitation of hybrid models in the current form is the time required for model training, which involves solving large systems of differential equations for each iteration of model optimization. For complex cases, this requires multiple CPUs to get decent computational time. To overcome this drawback more efficient algorithm and numerical techniques needs to be considered. With this, more complex FPMs can be used in the mechanistic framework of the hybrid models so as to further improve their performance.

ORCID

Massimo Morbidelli http://orcid.org/0000-0002-0112-414X

REFERENCES

Van Can, H. J. L. , Braake, H. A. B., Hellinga, C., Luyben, K. C. A. M., & Heijnen, J. J. (1997). An efficient model development strategy for bioprocesses based on neural networks in macroscopic balances. *Biotechnology and Bioengineering*, *54*(6), 550–566.

Creasy, A., Barker, G., Yao, Y., & Carta, G. (2015). Systematic interpolation method predicts protein chromatographic elution from batch isotherm data without a detailed mechanistic isotherm model. *Biotechnology Journal*, *10*(9), 1400–1411. https://doi.org/10.1002/biot.201500089

Food and Drug Administration (FDA). (2004 September). *Guidance for industry PAT: A framework for innovative pharmaceutical development. Manufacturing, and Quality Assurance*. FDA Official Document (pp. 16). Retrieved from https://www.fda.gov/CDER/guidance/6419fnl.pdf

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

Folch-Fortuny, A., Villaverde, A. F., Ferrer, A., & Banga, J. R. (2015). Enabling network inference methods to handle missing data and outliers. *BMC Bioinformatics*, *16*(1), 283. https://doi.org/10.1186/s12859-015-0717-7

Georgieva, P., Feyo de Azevedo, S., Gonçalves, M. J., & Ho, P. (2003). Modeling of sugar crystallization through knowledge integration. *Engineering in Life Sciences*, *3*(3), 146–153. https://doi.org/10.1002/elsc.200390019

Goudar, C. T., Joeris, K., Konstantinov, K. B., & Piret, J. M. (2005). Logistic equations effectively model mammalian cell batch and fed-batch kinetics by logically constraining the fit. *Biotechnology Progress*, *21*(4), 1109–1118. https://doi.org/10.1021/bp050018j

Hastie, T., Tibsharani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and predction*. Springer Series in Statistics (27, pp. 83–85). New York, NY: The Mathematical Intelligencer. https://doi.org/10.1007/b94608

Hu, G., Mao, Z., He, D., & Yang, F. (2011). Hybrid modeling for the prediction of leaching rate in leaching process based on negative correlation learning bagging ensemble algorithm. *Computers and Chemical Engineering*, *35*(12), 2611–2617. https://doi.org/10.1016/j.compchemeng.2011.02.012

Kroll, P., Hofer, A., Ulonska, S., Kager, J., & Herwig, C. (2017). Model-based methods in the biopharmaceutical process lifecycle. *Pharmaceutical Research*, *34*(12), 2596–2613. https://doi.org/10.1007/s11095-017-2308-y

Mercier, S. M., Diepenbroek, B., Wijffels, R. H., & Streefland, M. (2014). Multivariate PAT solutions for biopharmaceutical cultivation: Current progress and limitations. *Trends in Biotechnology*, *32*(6), 329–336. https://doi.org/10.1016/j.tibtech.2014.03.008

Montague, G. A., Glassey, J., Ignova, M., Paul, G. C., Kent, C. A., Thomas, C. R., … Ward, A. C. (2010). Hybrid modelling for on-line penicillin fermentation optimisation. IFAC Proceedings Volumes (*Vol. 35*). Retrieved from https://doi.org/10.3182/20020721-6-es-1901.01375

Nagrath, D., Messac, A., Bequette, B. W., & Cramer, S. M. (2004). A hybrid model framework for the optimization of preparative chromatographic processes. *Biotechnology Progress*, *20*(1), 162–178. https://doi.org/10.1021/bp034026g

Oliveira, R. (2004). Combining first principles modelling and artificial neural networks: A general framework. *Computers & Chemical Engineering*, *28*(5), 755–766. https://doi.org/10.1016/j.compchemeng.2004.02.014

Psichogios, D. (1992). A hybrid neural network—First principles approach to process modeling. *AIChE Journal*, *11*(2), 337–346. https://doi.org/10.1016/S0893-6080(98)00005-7

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

Schubert, J., Simutis, R., Dors, M., Havlik, I., & Lubbert, A. (1994a). Hybrid modeling of yeast production processes—Combination of a-priori knowledge on different levels of sophistication. *Chemical Engineering & Technology*, *17*(1), 10–20.

Schubert, J., Simutis, R., Dors, M., Havlik, I., & Lübbert, A. (1994b). Bioprocess optimization and control: Application of hybrid modelling. *Journal of Biotechnology*, *35*(1), 51–68. https://doi.org/10.1016/0168-1656(94)90189-9

Simon, L. L., Pataki, H., Marosi, G., Meemken, F., Hungerbühler, K., Baiker, A., … Chiu, M.-S. (2015). Assessment of recent process analytical technology (PAT) trends: A multiauthor review. *Organic Process*

Research and Development, *19*(1), 3–62. https://doi.org/10.1021/op500261y

Sokolov, M., Soos, M., Neunstoecklin, B., Morbidelli, M., Butté, A., Leardi, R., … Broly, H. (2015). Fingerprint detection and process prediction by multivariate analysis of fed-batch monoclonal antibody cell culture data. *Biotechnology Progress*, *31*(6), 1633–1644. https://doi.org/10.1002/btpr.2174

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

Tie Jong, S. (1993). *SIMPLS: An alternative approach to partial least squares regression*. (*Vol. 18*, pp. 251–263). Elsevier Science Publishers B.V. Retrieved from https://doi.org/10.1016/0169-7439(93)85002-X

Teixeira, A. P., Alves, C., Alves, P. M., Carrondo, M. J., & Oliveira, R. (2007). Hybrid elementary flux analysis/nonparametric modeling: Application for bioprocess control. *BMC Bioinformatics*, *8*(1), 30. https://doi.org/10.1186/1471-2105-8-30

Teixeira, A. P., Clemente, J. J., Cunha, A. E., Carrondo, M. J. T., & Oliveira, R. (2006). Bioprocess iterative batch-to-batch optimization based on hybrid parametric/nonparametric models. *Biotechnology Progress*, *22*(1), 247–258. https://doi.org/10.1021/bp0502328

Teixeira, A. P., Cunha, A. E., Clemente, J. J., Moreira, J. L., Cruz, H. J., Alves, P. M., … Oliveira, R. (2005). Modelling and optimization of a recombinant BHK-21 cultivation process using hybrid grey-box systems. *Journal of Biotechnology*, *118*(3), 290–303. https://doi.org/10.1016/J.JBIOTEC.2005.04.024

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

Thompson, M. L., & Kramer, M. A. (1994). Modeling chemical processes using prior knowledge and neural networks. *AIChE Journal*, *40*(8), 1328–1340. https://doi.org/10.1002/aic.690400806

Tian, Y., Zhang, J., & Morris, J. (2001). Modeling and optimal control of a batch polymerization reactor using a hybrid stacked recurrent neural network model. *Industrial & Engineering Chemistry Research*, *40*(21), 4525–4535. https://doi.org/10.1021/ie0010565

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. (2011). A novel identification method for hybrid (N)PLS dynamical systems with application to bioprocesses. *Expert Systems with Applications*, *38*(9), 10862–10874. https://doi.org/10.1016/j.eswa.2011.02.117

Von Stosch, M., Hamelink, J.-M., & Oliveira, R. (2016). Hybrid modeling as a QbD/PAT tool in process development: An industrial *E. coli* case study. *Bioprocess and Biosystems Engineering*, *39*(5), 773–784. https://doi.org/10.1007/s00449-016-1557-1

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

Yang, A., Martin, E., & Morris, J. (2011). Identification of semi-parametric hybrid process models. *Computers and Chemical Engineering*, *35*(1), 63–70. https://doi.org/10.1016/j.compchemeng.2010.05.002

Zander, B. H., & Dittmeyer, R. (1999). Dynamic modeling of chemical reaction systems with neural networks and hybrid models. *Chemical Engineering & Technology*, *21*(7), 571–574.

Zhang, J., Mao, Z. Z., Jia, R. D., & He, D. K. (2015). Real-time optimization based on a serial hybrid model for gold cyanidation leaching process. *Minerals Engineering*, *70*, 250–263. https://doi.org/10.1016/j.mineng.2014.09.021

Zorzetto, L. F. M., Filho, R. M., & Wolf-Maciel, M. R. (2000). Processing modelling development through artificial neural networks and hybrid models. *Computers & Chemical Engineering*, *24*(2–7), 1355–1360. https://doi.org/10.1016/S0098-1354(00)00419-1

Zorzetto, L. F. M., & Wilson, J. A. (2003). Monitoring bioprocesses using hybrid models and an extended Kalman filter. *Computers & Chemical Engineering*, *20*(96), S689–S694. https://doi.org/10.1016/0098-1354(96)00124-x

How to cite this article: Narayanan H, Sokolov M, Morbidelli M, Butté A. A new generation of predictive models: The added value of hybrid models for manufacturing processes of therapeutic proteins. *Biotechnology and Bioengineering*. 2019;116:2540–2549. https://doi.org/10.1002/bit.27097