Received: 12 September 2023 | Revised: 27 February 2024 | Accepted: 29 February 2024

DOI: 10.1002/biot.202300473

Biotechnology Journal

RESEARCH ARTICLE

An innovative hybrid modeling approach for simultaneous prediction of cell culture process dynamics and product quality

Jakub Polak[1] | Zhuangrong Huang[2] | Michael Sokolov[1] | Moritz von Stosch[1] | Alessandro Butté[1] | C. Eric Hodgman[2] | Michael Borys[2] | Anurag Khetan[2]

[1]DataHow AG, Zurich, Switzerland

[2]Biologics Development, Global Product Development and Supply, Bristol Myers Squibb, Devens, Massachusetts, USA

Correspondence
C. Eric Hodgman, Biologics Development, Global Product Development and Supply, Bristol Myers Squibb, 38 Jackson Rd, Devens, MA 01434, USA.
Email: eric.hodgman@bms.com

Alessandro Butté, DataHow AG, Hagenholzstrasse 111, 8050 Zürich, Switzerland.
Email: a.butte@datahow.ch

Abstract

The use of hybrid models is extensively described in the literature to predict the process evolution in cell cultures. These models combine mechanistic and machine learning methods, allowing the prediction of complex process behavior, in the presence of many process variables, without the need to collect a large amount of data. Hybrid models cannot be directly used to predict final product critical quality attributes, or CQAs, because they are usually measured only at the end of the process, and more mechanistic knowledge is needed for many classes of CQAs. The historical models can instead predict the CQAs better; however, they cannot directly relate manipulated process parameters to final CQAs, as they require knowledge of the process evolution. In this work, we propose an innovative modeling approach based on combining a hybrid propagation model with a historical data-driven model, that is, the combined hybrid model, for simultaneous prediction of full process dynamics and CQAs. The performance of the combined hybrid model was evaluated on an industrial dataset and compared to classical black-box models, which directly relate manipulated process parameters to CQAs. The proposed combined hybrid model outperforms the black-box model by 33% on average in predicting the CQAs while requiring only around half of the data for model training to match performance. Thus, in terms of model accuracy and experimental costs, the combined hybrid model in this study provides a promising platform for process optimization applications.

KEYWORDS
CHO cell, hybrid model, machine learning, product quality

1 | INTRODUCTION

The Process Analytic Technology (PAT) initiative, published in 2004 by the US Food and Drug Administration,[1] opened the door for mathematical model-supported process development and manufacturing. Subsequent developments, such as the Quality by Design (QbD) paradigm being adopted in the ICH guidance documents, further promoted the role of mathematical modeling alongside risk-based approaches for decision-making.[2–5] Using a risk-based approach, the process parameters that might critically impact product critical quality attributes (CQAs) are defined as critical process parameters (CPPs). The definition of the precise relation between CPPs and CQAs can be very challenging, considering the complex nature of manufacturing processes, especially cell culture-based processes. This is further complicated by the need to identify the process design space, that is, the CPPs' space, where operating the process without jeopardizing the

Abbreviations: BBM, black box model; CHM, combined hybrid model; DHM, discrete hybrid model; DoE, design of experiments; HM, historical model; PLS, partial least squares; PM, propagation model.

Jakub Polak and Zhuangrong Huang contributed equally to this work.

CQAs is possible. In this context, process modeling becomes of great value in supporting such activities, particularly when considering the need to reduce process development time and costs simultaneously.

Multiple factors, such as process characteristics and goals, data availability, and model characteristics and expectations with respect to precision, robustness, and extrapolation, influence the choice of the process model. Black box models, such as multiple linear regression models, are often used due to their simplicity, as they are purely data-driven and do not require any specific prior knowledge of the process. These models can also generate experimental designs simply and effectively (e.g., factorial designs), which can return helpful and practical information on variable importance and interactions. However, the specific knowledge generated by such models is limited, as they might require a significant amount of data to reach good prediction accuracy, especially in the presence of many process variables. Additionally, due to their structural simplicity, black box models can only provide reasonable approximations of the process behavior in limited design space.

In contrast, mechanistic models are derived from a first-principle process description. They might require a minimal amount of training data when (1) well-defined prior knowledge is available (i.e., a set of differential equations are available to describe the targeted process), and (2) it is possible to design ad-hoc experiments to estimate the physical parameters of the model. The possibility of generating process insights from such models is very high, as is the possibility of extrapolation.[6] There are several examples of mechanistic models applied to the modeling of CQAs for therapeutic proteins. Glycosylation has been modeled in detail, allowing a better understanding of glycoform biosynthesis.[7–9] However, the amount of prior knowledge and the need to create new knowledge for more complex drug targets strongly limits this approach.

Hybrid models represent an attractive compromise between these two approaches.[6,10,11] According to this new paradigm, it is possible to define a sufficiently general description of the process using a proper combination of first-principle models (e.g., by describing process mass balances and main kinetic characteristics) and data-driven models to compensate for the lack of knowledge on the process dependency upon all the process parameters. Although the use of hybrid models for bioprocesses is well documented,[12–15] they have been applied only to the description of the process evolution in time, that is, for quantities like cell density and metabolite concentrations, as a function of the choice of the process parameters. Little work is available to describe the impact of CPPs on the final values of the CQAs.

In a previous work,[16] we suggested a two-step approach to create models for CQAs. Firstly, a hybrid model is used to describe the process evolution as a function of the process parameters. Secondly, this evolution is fed into a historical model to explain how such evolution is impacting the final CQA values. The use of historical models for product CQAs has been widely documented[17,18] and is common in the industry, for instance, with the use of multivariate (e.g., Partial Least Squares, PLS) models. These models can provide detailed descriptions of CQAs without requiring prior knowledge of each CQA and without needing a large reference dataset. In addition, it is possible to have a detailed analysis of the impact of process parameters and conditions to support a specific understanding of critical process conditions.[19]

In this work, we investigated the combination of hybrid models to describe process evolution with historical models to describe product CQAs. This combined model is referred to as the Combined Hybrid Model which directly connects the process parameters to the final product quality by capturing the full process dynamics. To do this, we analyzed the performance of hybrid and historical models alone in predicting process evolution and CQAs, respectively. Then, we analyzed the performance of the combined hybrid model: first, how prediction errors are propagated from the propagation to the historical model; second, how the performance of the combined hybrid model is impacted by the use of different numbers of experiments to train the model, in terms of model precision and robustness; and third, how the model can support a detailed analysis of CQA behavior within the process design space. Specifically, the performance of the combined hybrid model is compared to a fully data-driven (or black box) model, whose usage is considered industrial practice to create a basic knowledge of how CPPs impact CQAs. A typical set of experiments was used in this study, for which both the full process evolution in time and multiple CQAs of different natures were measured. The current work thus represents one of the first attempts to characterize the dependence of product CQAs from process parameters on a broad set of industrial data using hybrid models.

2 | MATERIAL AND METHODS

2.1 | Variable nomenclature

Throughout this work, we refer to variables according to the following nomenclature, as adopted in[20]: independent variables expressing process conditions that are constant throughout the run are referred to as Z variables; observed uncontrolled (dependent) variables as X (also referred to as process state or response); and controlled independent variables as W. Among the latter, we indicate flows or feeds as F variables, as they actively contribute to the mass balance of some X variables. Finally, variables typically only measured once (e.g., product CQAs) near the end of the experiment are referred to as Y variables. These block variables are used to showcase the generalizability of the methodology summarized within this manuscript.

2.1.1 | Fed-batch dataset for modeling

As shown in Table 1, the dataset for this work comprises 48 fed-batch experimental conditions. Four independent custom designs of experiments (DoEs) were performed during early process development. Each DoE was used to study the combined effects of several parameters, as detailed in Table 1. Overall, the dataset investigated the impact of MSX (methionine sulfoximine), basal media, feed media, feed volume, inoculation density, temperature shifts, and pH settings in production

TABLE 1 Studied parameters for the four different DoEs.

DoE	Batches	Studied parameters
DoE 1	12 × 5L	MSX levels in seed train, Temperature shift, Feed medium, Feed volume
DoE 2	12 × 5L	Inoculation density, Temperature shift, pH setting, Feed volume
DoE 3	12 × 5L	MSX levels in seed train, Basal medium, Feed medium, Temperature shifts
DoE 4	12 × 5L	Inoculation density, Basal medium, Feed medium, Feed volume

cultures. The variation of the studied parameters can be seen in Figure S1 in the support material. Similarly, the resulting measured variation in the target CQAs space can be seen in Figure S2. Note that traditional DOEs are not necessary to train the model, and any experiments with variation in process parameters and product quality attributes can be used for modeling.

2.1.2 | Cell line and cell culture processes

A CHOZN GS cell line was used for the expression of a proprietary IgG1 mAb. Proprietary, chemically defined seed, basal, and feed media were used in this study. CHO seed culture was started from the thaw of a working cell bank vial. All seed cultures were performed using shake flasks (Corning Life Sciences) in a humidified incubator (Climo-Shaker, Kuhner). Cells were passaged every 3 or 4 days prior to fed-batch production. The fed-batch processes were performed in 5-L glass bioreactors (Sartorius) to investigate the process parameters described in Table 1. Additional glucose was added as needed to maintain its concentration at a fixed level. The bioreactor pH was controlled by the addition of CO2 gas or 1 M Na_2CO_3 base as needed. Dissolved oxygen (DO) was maintained by oxygen sparging.

2.1.3 | In-process cell culture and quality attribute analysis

Off-line pH, pCO^2, and pO^2 were measured using a Bioprofile pHOx analyzer (Nova Biomedical). Viability cell density (VCD) and cell viability were quantified off-line by a Vi-CELL XR automatic cell counter (Beckman Coulter) using trypan blue dye exclusion. Metabolites, including glucose, glutamine, glutamate, lactate, and ammonia, were quantified using a CEDEX Bio HT analyzer (Roche). Titer was determined using a Protein A titer assay after centrifugation of supernatant samples at 1000 × *g* for 5 min. The maximum titer on day 14 among all bioreactors was set as 100% to normalize the titer within this study. In addition, Protein A purified samples were used for analyzing product quality attributes, including impurities, charge variants, intact mass, size variants, and N-glycan profiling. All assays were developed by the analytical team at Bristol Myers Squibb.

To summarize process dataset descriptions, each of the variables was assigned to their corresponding blocks:

- Z variables: MSX, N-1 stage total cell count (Nm1VCD), N stage inoculation density (InocVCD), glucose concentration at inoculation (InocGlc), first temperature shift time (TempShift1), second temperature shift value (TempValue2) and time (TempShift2), third temperature shift value (TempValue3), media (i.e., BasalMedium, FeedMedium, and FeedMedium2, which are all boolean variables), pH settings (pH1low, pH2high) and feed volume and concentration.
- X variables: viable cell density (VCD), glucose (Glc), glutamine (Gln), glutamate (Glu), lactate (Lac), ammonia (Amm), cell viability (Viabi), and titer.
- W variables: online pH, pCO^2, DO, and temperature (T).
- F variables: continuous feed rate and bolus feed masses of Glc, conditional on feeding level.
- Y variables: product quality attributes, including impurity levels, charge variants species (main, acidic, and basic peaks), intact mass (Intact, MonoClipped, and DiClipped species), total low molecule weight (LMW), and N-glycan profiling.

2.2 | Models for process evolution (X variables)

In this work, we utilize the hybrid model of Narayanan et al.[20] to capture the evolution of the process. As Narayanan's work has demonstrated, hybrid models are a better option than purely data-driven or mechanistic models for predicting cell culture, and these models are essential for the proposed modeling approach. Below, we introduce our implementation of the discrete hybrid model with Gaussian Processes.

2.2.1 | Propagation model (PM): Discrete hybrid model (DHM)

A discrete hybrid model was used to characterize the time evolution of the X variables as a function of the process conditions (Z), the initial condition of the X variables, and the dynamic process control variables (W and F). Due to its discrete nature, the model predicts the evolution of the process (X) variables at discrete intervals only, that is, in correspondence with the process measurements.

First, a generic mass balance is written for all species:

$$\frac{dc \cdot V}{dt} = R(s) \cdot V + u_f \tag{1}$$

where c is a vector of concentrations (e.g., VCD); *t* time; *s* a vector defining the process states including the value of all Z, X, W, and F variables at time *t*; *V* the culture volume; and u_f the continuous mass feed rate or amount of bolus of F variables. The term *R*(*s*) represents the total rate of production (or consumption, when negative) of a species as a function of the process state *s* at time *t*. Using the forward-difference formula for the concentration time derivative *dc/dt* in Equation 1, one

obtains:

$$\left.\frac{dc \cdot V}{dt}\right|_{t_i} = V \cdot \left.\frac{dc}{dt}\right|_{t_i} + c \cdot \left.\frac{dV}{dt}\right|_{t_i} \cong V \cdot \frac{c(t_{i+1}) - c(t_i)}{t_{i+1} - t_i} + c \cdot \frac{dV}{dt}$$
$$= R(s) \cdot V + u_f \quad (2)$$

where $c(t_i)$ is the concentration of the species measured at time t_i. Rearranging Equation 2, one obtains the explicit definition of the rate of accumulation at time t_i as a function of the process state s:

$$R(s) = \frac{c(t_{i+1}) - c(t_i)}{t_{i+1} - t_i} - \frac{1}{V} \cdot \left[u_f - c(t_i) \cdot \frac{dV}{dt}\right] \quad (3)$$

In the proposed model, the rate of accumulation of each species is modeled through a Gaussian Process, that is, $R(s) \approx GP(s)$. In other terms, the Gaussian Process model infers $R(s)$ from the knowledge of the state s of the process at each time t_i. When predicting the process evolution from a generic state s, the following expression is then used:

$$c(t_{i+1}) \approx c(t_i) + \left(GP(s) \cdot V + u_f - c(t_i) \cdot \frac{dV}{dt}\right) \cdot \frac{t_{i+1} - t_i}{V} \quad (4)$$

Each prediction timestep of the model only depends on the previous process state. As such, this model can also be referred to as propagation model, as prediction occurs by propagating the process states from the initial state in finite steps, defined by the time steps t_i at which the measurements of the X variables are available. In theory, a machine learning model other than Gaussian Processes can be used to approximate the reaction rates, such as Neural Networks. However, we found that Gaussian Processes are better suited for situations with fewer training experiments and are therefore used in this work. In Section 3.1, we also compare the performance of this Discrete Hybrid Model with Gaussian Processes to a Continuous Hybrid Model with Neural Networks.

2.2.2 | Training/test split and model ensembling

To assess the performance of the models, the experiments were divided into a fixed training and test set with a ratio of 3:1 (i.e., the training set comprises 36 experiments and the test set 12 experiments). The test set was selected from the performed custom DoE so that experiments are representatively sampled across the entire design space. Our previous study described the detailed methodology.[21] In short, similar experiments are clustered together once the experiments have been transformed using a Principle Component Analysis (PCA). Each cluster is sampled randomly to obtain a representative training set of the desired size. This test selection methodology was used due to the custom structure of the DoE performed and the availability of the data.

A single propagation model contains 20 submodels, where each is trained on 75% randomly chosen experiments from the training set (i.e., 24 experiments). In this work, model ensembling in prediction is obtained by simple averaging. Firstly, all models in the ensemble are used to estimate the value of GP in Equation 4 as a function of the estimated process state s at that point in time. Secondly, the final value of $GP(s)$ to be used in Equation 4 to estimate the value of $c(t_{i+1})$ is obtained by averaging the values from each model.

2.3 | Models for product quality (Y variables)

Two types of models are used in this work to predict the final process state and the final product CQAs (i.e., Y variables): the black box model (BBM) and the historical model (HM). Both models are considered state-of-the-art in the industry. The BBM directly correlates the manipulated process parameters (Z variables) to the corresponding CQAs (Y variables) and, therefore, can be used to design or predict new experiments. The HM correlates the entire evolution of the process in time (X variables) to the corresponding Y variables. The HM is expected to be more precise, as it embeds all the process history. However, it cannot be used to predict new experiments, as it requires process history knowledge. To enable the ability to predict new experiments, we combine the PM (i.e., discrete hybrid model) with the HM to create the Combined Hybrid Model (CHM), which is the object of this work's investigation. The performance of CHM is evaluated by comparing it to the BBM.

2.3.1 | Black box model (BBM)

A simple but effective approach to process modeling is to model Y variables by using only process conditions (Z variables). Such model is here referred to as the Black Box Model. Due to its simplicity, this approach is frequently adopted in combination with response surface methodologies. The Y variables are treated as independent response variables with a specific model for each quality attribute. Therefore, the model simply estimates CQAs under the association $Y = f(Z)$. For such a model, the entire design space can be explored, and new experimental conditions can be evaluated. The major drawback of the model is that it does not consider process evolution, and it produces estimates only for a one-time point, typically the end of cultivation.

This work used a multiple linear regression (MLR) model, including interaction and quadratic terms. As explanatory variables, all manipulated variables (i.e., all Z variables) used in the definition of the original DoEs were applied as BBM model inputs. Note that the number of potential explanatory variables when using interactions and quadratic terms largely exceeds the number of experiments (e.g., in this work, 15 main factors + 105 interactions + 15 quadratic terms). To train the model, a classical stepwise regression algorithm was implemented to add and remove statistically relevant terms iteratively until the model with the largest coefficient of determination (R^2) is found.

2.3.2 | Historical model (HM): PLS model

In the HM, the full evolution of the cell culture process variables is used as an input to model the final process properties, that is, the Y

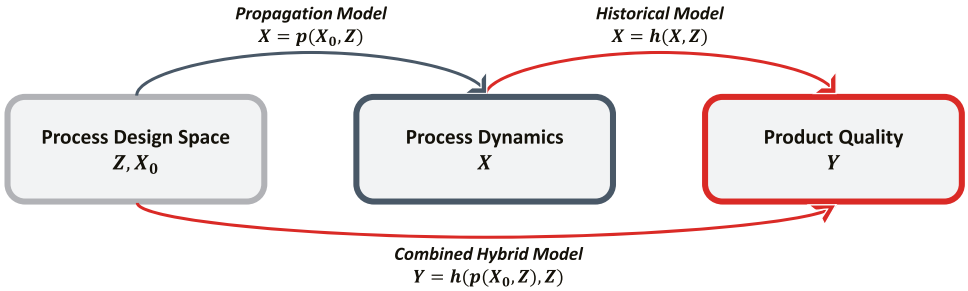

FIGURE 1 The depicted schema for the two-step predictive methodology to connect process design (Z, X_0) to final product quality (Y) based on the process dynamics (X) prediction. The first step consists of the propagation model p (i.e., discrete hybrid model) that takes the process initial conditions (X_0) and process design space (Z) as inputs and autoregressively propagates them to obtain full process dynamics (X). The second step consists of the historical model h (i.e., PLS), which takes the full process dynamics (X) as an input and predicts the product quality attributes (Y) as outputs. A combined hybrid model linking process design space to product quality is obtained by combining these two models in a hierarchical way. In practice, the predicted full process dynamics from the propagation model are used as inputs to the historical model.

variables. This work used a standard partial least squares (PLS) model for this scope. Specifically, the model has the following structure:

$$Y = h\left(\vec{X}, \vec{Z}, \vec{W}\right) \tag{5}$$

where $\vec{X}, \vec{Z}$, and $\vec{W}$ are the batch-wise unfolded representation of all process variables. In other words, according to this scheme, all data from one experiment are unfolded in a single matrix row: this comprises all the Z variable values for that experiment, and all the values from the X and W variables at all sampling times. For each Y variable, a single PLS model is created, using a 5-fold cross-validation (after random permutation of the experiments) to define the optimal number of latent variables, and the root mean squared error (RMSE) in cross-validation as criterion (see Section 2.4 for definition).

The choice of PLS has a two-fold justification: (1) this type of model has been widely used to model product CQAs in the industry; (2) a PLS model offers the possibility of optimally handling collinear variables: for example, the X and W values measured at consecutive days are likely to be correlated to each other.

2.3.3 | Combined hybrid model (CHM)

We obtain the Combined Hybrid Model by combining propagation and historical models, as shown in Figure 1. With this CHM model, the user can predict how the final product quality attributes (or CQAs) are impacted by the choice of the process parameters in two steps: the process evolution is first predicted by the PM, which is then used as an input to predict the Y variables by the HM. Therefore, CQAs can be predicted via the CHM as:

$$Y = h\left(\hat{\vec{X}}, Z, \vec{W}\right) = h\left(p\left(X_0, \vec{Z}, W\right), Z, \vec{W}\right) \tag{6}$$

where p represents the propagation model and the h historical model; $\hat{\vec{X}}$ indicates the batch-wise unfolding of the X variable estimated by the PM. The primary mechanism in the CHM is linked to the hybrid or propagation model, which predicts the process evolution. Without this full process evolution predicted from the process design space, we could not apply the historical model to predict the product CQAs. The main distinction between CHM and BBM is that the dynamic process variables are measured and worked as a bridge in CHM to link the manipulated process variables to product CQAs, as seen in Figure 1. It is essential to note that the HM is trained independently by using the measured values of the X variables and not those estimated by the PM. This procedure enables process simulation for untested process design conditions and optimization toward product quality specifications.

2.4 | Model evaluation metrics

The primary metric used to evaluate model performance is the root mean squared error (RMSE) in prediction computed for the experiments in the test set.

$$RMSE_i = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

with n the number of residuals, y_i the measured values, and $\hat{y}_i$ the predicted ones for the i-th X variable. To compare errors in prediction across different variables, we normalized the metric with respect to the variability of each variable, for which we use the standard deviation σ_i on all experiments as a proxy. The relative RMSE is calculated as follows: $RMSE_i / \sigma_i$. The predictions of a model can be considered unreliable if the relative RMSE is above 1, as this corresponds to a value of the coefficient of determination R^2 smaller than zero.

Another important metric is the replicate error, which represents the process variation with respect to "identical" design conditions. If the experiment is replicated, it is the amount by which an aleatoric uncertainty causes a change in either the X or Y variables. This provides a lower bound to the model error of each variable.

3 | RESULTS AND DISCUSSION

In Section 3.1, we first analyzed the performance of the PM as a prerequisite to the success of the CHM. This model combines mass balances on all species with Gaussian Processes, which describe each species' rate of change as a function of the process state. In Section 3.2, we proceeded to describe the HM using the widely used PLS approach. In Section 3.3, the final CHM was generated by combining the PM and HM and compared to the BBM in terms of accuracy and robustness. Section 3.4 discussed how the number of experiments used for model training impacts model precision and robustness. We further analyzed the prediction capabilities of all models so far described in Section 3.5 and qualitatively discussed how such models can differ in predicting and explaining the dependences of the product CQAs within the design space, as such type of analysis is of paramount importance in defining the process design space according to the QbD paradigm.

3.1 | Propagation model (PM)

Discrete propagation models fall in the broader category of Markov chain models. In such models, the knowledge of the process state at a certain point in time is used to estimate the probability that the process evolves into a new state. Once inserted into the frame of mass balance on each process species, the mass balance adds robustness to the model, for instance, in the case of the depletion of some species. In this context, the PM predicts the change in the state of the process, that is, ΔX (see Section 2.2.1). Optimal results are obtained when the predicted X variables are measured at regular (e.g., daily) intervals, as is the case in this study.

The performance of the PM is shown in Figures 2 and 3. In Figure 2, the predicted values of all the dependent (X) variables are plotted versus the corresponding measured values for both training (orange dots) and test (blue diamonds) experiments at all process times. The plot diagonal (dashed black line) represents the perfect model prediction (i.e., the predicted values are equal to the measured values). A quite satisfactory correspondence between predicted and measured values can generally be observed. In addition, the predicted test values are similarly scattered around the diagonal as the training values, indicating that the model prediction does not lose significant precision. However, two variables (glucose and lactate) seem to have larger prediction errors than the others. It is interesting to note that the largest deviations from the measurement are observed in glucose for the training data. In this case, some measurements are significantly underpredicted, and all originated from the same experiment (Exp. BRX3-V204). The accuracy of the model is clearly worse than other variables, for instance, VCD, as indicated by a broader scattering around the diagonal. However, the model retains statistical significance, being accurate for both low and high glucose measurements, and it does not show a significant visual deterioration of its predictive capabilities with respect to the training set. To explain this behavior, it must be noted that glucose prediction is rather sensitive to input errors due to the structure of the hybrid model. In fact, a bolus feed or a change in continuous feed frequently takes place at a different time than the reported sampling time, which can cause an offset in glucose prediction.

In the case of lactate, a very good agreement in prediction was observed for the training set, while the prediction of the test set is significantly worse. Nonetheless, there is still a satisfactory agreement between models and measurements, and the largest deviations are often confined to a few single experiments. To a lesser degree, a similar observation can be made for both glutamine and ammonia. The fact that the largest source of error can be traced back to single experiments which is consistent with the observation that all such species are highly interlinked and prone to suffer propagation errors. This effect can be well observed in Figure S3 in the support material, where the measured time series profiles for the main X variables (blue curves) are compared to the model predictions in time (orange curves) for the test experiments only. For instance, a significant deviation in the glucose profile is seen in replicate experiments BRX1-V203 and BRX1-V204 on day 11. While the data suggest a mild decrease in glucose concentration with respect to day 10, the model predicts a very large increase, due to the presence of a large bolus injection on the same day. Although it is not possible to arrive at any conclusive evidence, the deviation is likely caused by an error in the data, given that the model predicted the evolution of glucose very precisely for the previous 10 days. Most likely, as mentioned above, the bolus feed took place later in time, thus causing a smaller change in glucose and the offset measured on day 11. Note that, on the same day, the glutamine, glutamate, and lactate predictions are still rather precise. Thus, it is unlikely that the significant prediction error on glucose observed on day 11 is caused by these variables. On the contrary, the prediction on these two variables starts accumulating errors from this point on, probably due to the mistakes in glucose. Similar patterns can be observed in most of the other experiments. For instance, a significant glucose deviation in BRX2-V207 observed on day 6 is causing a similar deviation in the lactate profile. This analysis confirms the strong interlink among these species. It suggests that errors in the glucose prediction are probably due to data misalignment, which in turn causes the progressive deviation in both glutamine and lactate.

Figure 3 shows the calculated relative RMSE for each X variable for both the training (orange bars) and test sets (blue bars). It can be noted that the relative RMSE in the test set for titer is always smaller than 0.45 and can be as low as 0.10. There is no substantial difference between the error in training and test for glucose, supporting the hypothesis that this is mainly driven by offsets in glucose feed times. As the effect of feeds is directly accounted for in the mechanistic part of the model, any error in the data (such as feeding time and feeding amount) is directly transmitted to the model prediction, both in the training and the test set. On the other hand, feeds are not primarily impacting other variables because a much smaller amount was fed compared to glucose, for example, glutamine, glutamate, and lactate, and the Gaussian Process model can learn their evolution: with training, this produces very small prediction errors; however, once in prediction, any deviation occurring to glucose is propagated to

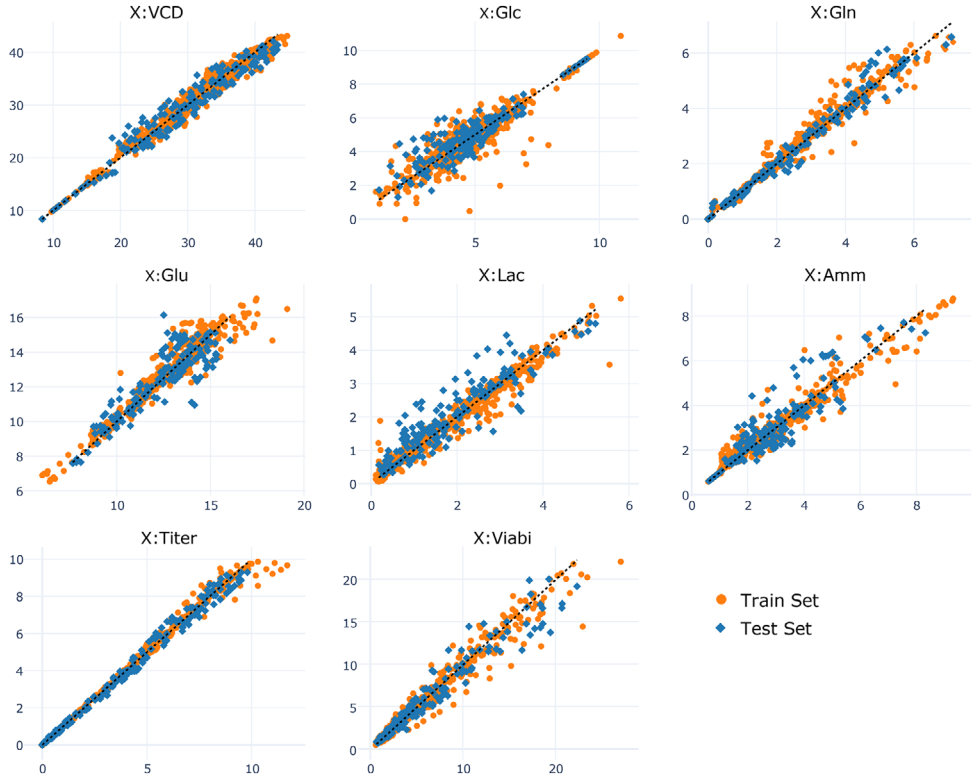

FIGURE 2 Predicted (x-axis) versus measured (y-axis) values for different X variables. Orange dots are observations in training experiments; blue diamonds are observations in test experiments; a dashed black line is a fitting line.

these variables, which are naturally sensitive (and, thus, correlated) to deviations in glucose.

For comparison, the same errors were calculated using the continuous hybrid model described by Narayanan et al.[20] and also utilized by Cruz et al.[22] This model embedded a neural network to calculate the specific rate of consumption/production of each species and is integrated in time without the need for any discretization. The prediction on glucose in the test experiments is slightly improved (relative RMSE of 0.34 vs. 0.42 for the PM). Consequently, the glutamate, lactate, and ammonia predictions are also slightly better (0.40, 0.31, and 0.37, respectively). However, besides this improvement, the discretization included in the PM does not substantially change the quality of the PM. Thus, the discrete models still represent a valid and quick alternative to the continuous models. This is true for the data type discussed in this work, where each variable is sampled daily. As mentioned in Narayanan et al.,[20] continuous models have the advantage of being

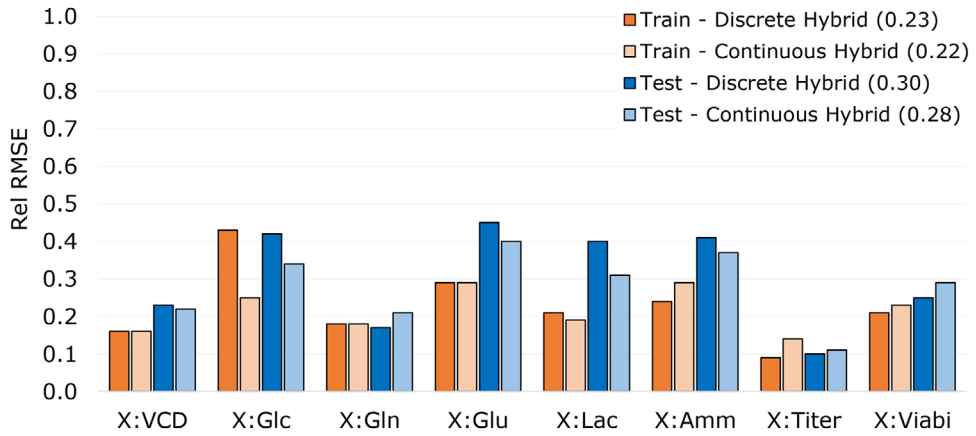

FIGURE 3 Calculated relative RMSE for different X variables using the PM. Orange bars are training experiments; blue bars are test experiments. Dark bars correspond to the Discrete Hybrid Model, and light bars correspond to the Continuous Hybrid Model. On the top right, the average value of the relative RMSE values for all X variables is reported for both sets of experiments and models.

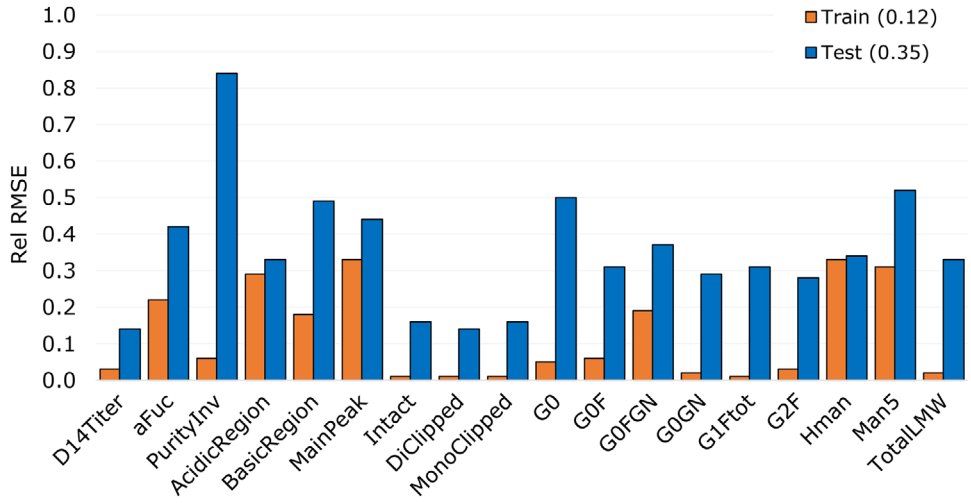

FIGURE 4 Calculate relative RMSE for different Y variables using the HM. Orange bars are train experiments; blue bars are test experiments. On the top right, the average value of the relative RMSE values for all Y variables is reported for both sets of experiments.

able to easily handle misaligned data, missing data, and variables sampled at very different rates (e.g., from on-/at-line measurements) and hence are expected to return significantly better predictions.

To conclude, the PM analysis indicates that the model has excellent capabilities of predicting the evolution of the X variables in time, even though the operative conditions were changed broadly according to the custom DoE scheme. Thus, only the extreme values of the design variables were tested, while Gaussian Processes operate best when uniform designs are used.[23] Most of the deviations are traceable to errors in the glucose predictions, which are likely caused by data misalignment. Despite this, the PM has proven to be a robust tool to predict the evolution of the process, which is needed as an input to the HM, as will be discussed in the next section.

3.2 | Historical model (HM): The historical PLS model

The constant independent variables (Z) and the evolution in time of the X variables were used to estimate the final state of the process and, most importantly, the product CQAs. The rationale behind the HMs is that the final product quality is the outcome of the cumulated behavior of the process, which in turn is primarily related to the time evolution of the measured variables. Due to its properties, PLS regression (also known as PLS1 for single targets and as PLS2 for multiple targets) is often used for this purpose, being able to handle highly correlated variables (i.e., the measurements at different times), reduce the space to a few latent variables, and handle measurement noise. In this work, we created a PLS1 model for each target (Y) variable individually, although PLS2 models for highly correlated targets are known to be more robust.[19] As these models are prevalent in industrial practice, we simply describe the performance of such model applied to the data of this work, as its relevance will impact and be relevant to the description of the CHM.

Figure 4 shows the calculated relative RMSE on each Y variable for the training (orange bars) and test set (blue bars). The model regresses the data remarkably well and retains excellent prediction capabilities even on the test set, as confirmed by the average value of the relative RMSE across all Y variables (0.35). The only notable exception is represented by ImPurity, although well regressed in the training set, which shows no significant statistical relevance in the test set, with the relative RMSE being close to one (0.84). These results are even more important when accounting for the replicate error. This was calculated independently on all Y variables using the center points of the different DoEs, as shown in Figure 5 (gray bars). The average value of the replicate relative error is 0.14. For many variables, namely, AcidicRegion, MainPeak, Intact, DiClipped, and MonoClipped, this represents a significant portion of the relative RMSE observed in the test set. Thus, the HM accounts for a substantial portion of the observed and explainable variability and, therefore, represents a robust tool to predict product CQAs, provided that the evolution of the process variables is known. Finally, one could also argue that a more non-linear model than PLS could reduce the error in the prediction of the HM. However, the PLS model was chosen for being very common in the industry and for its robustness, especially when dealing with a limited number of experiments.[19]

3.3 | Combined hybrid model (CHM): Comparison with the black box model (BBM)

The combination of PM and HM, that is, the CHM, allows to directly link the final properties of the process and the product CQAs to the manipulated process parameters. Due to the nature of the two models discussed in the previous sections, such models can potentially handle very complex and nonlinear dependencies among the process parameters and CQAs. In this section, we evaluated this capability and the robustness of such models and compared the results of the CHM to

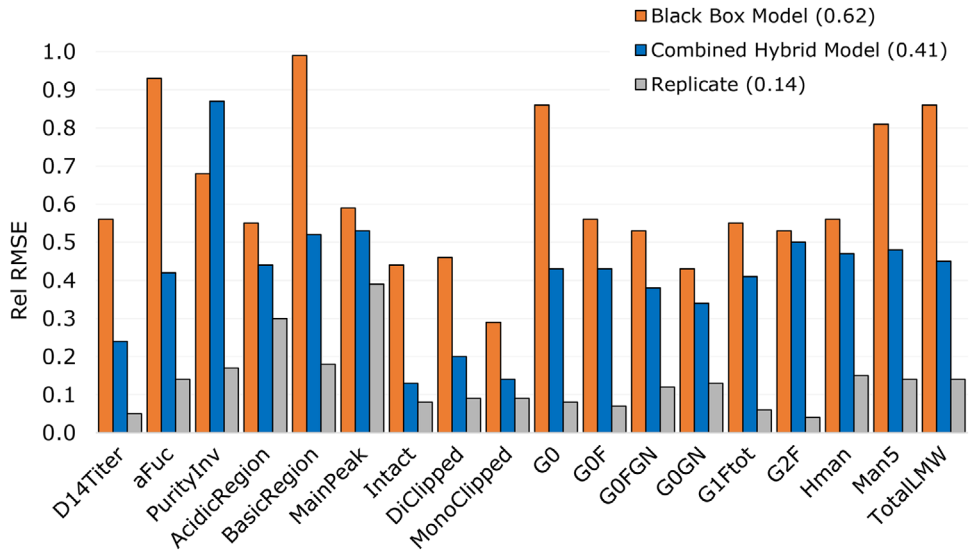

FIGURE 5 Relative RMSEs were calculated on the Y variables for the test set with the BBM (orange) and the CHM (blue). The replicate error (grey) calculated on a whole dataset is shown for comparison.

the BBM. The latter is based on a multiple linear regression model with interaction and quadratic terms. However, such a model is expected to show some precision limitations, compensated by good robustness.

Figure 5 compares the relative RMSE calculated for the CHM on all Y variables for the test set (blue bars) to that for the BBM (orange bars). As in the previous plots, all errors are standardized using the standard deviation, calculated on all data for each variable. In this figure, it is observed that, except for ImPurity, the relative RMSE of the BBM is much larger than the one obtained using the CHM despite the much larger functional complexity of the latter. To elucidate such a difference, we compared the mean of the relative RMSE across all variables: 0.62 for the BBM versus 0.41 for the CHM. Most importantly, by comparing the error to that of the HM shown in Figure 4, it can be observed that the two plots show similar errors, as also confirmed by the average value of the relative RMSE of the HM (0.35) in Figure 4, which is just slightly lower than that of the CHM model (0.41). This result suggests that the PM contributes little to the error in prediction, that is, there is little error propagation between the two models. For this reason, the combination of the continuous hybrid model (Figure 3) with the PLS model does not show significant improvement to the overall prediction error (not shown). It is worth highlighting that, as noted in Section 2.3.1, where the BBM is described, the number of experiments in the training set (36) is smaller than the potential number of interaction terms (105), given that 15 explanatory variables (Z) were used. Therefore, it is clear that, under such conditions, a linear regression model is heavily limited by the number of experiments. This will become even more evident when the training set's size is reduced in the next section. On the other hand, this adds a different perspective on the importance of the CHM, as both the PM and the HM model can be trained with a very small number of experiments.

Figure 6 plots the predicted values on all Y variables for the test experiments versus the corresponding measured values for the BBM (orange dots) and the CHM (blue diamonds). In the same figure, both sets were linearly regressed (orange and blue dashed lines, respectively), and these regression lines can be compared to the diagonal (black dashed line), representing the perfect fit. The more the regression line is close to the diagonal, the better the "goodness" of the prediction. For many variables, both regression lines lie remarkably close to the diagonal, indicating a very good prediction on the test set (e.g., Intact, MonoClipped, G0F, G0FGN, G0GN, G1Ftot). For some variables, the regression line for the CHM is clearly closer to the diagonal than that for the BBM (e.g., D14Titer, DiClipped, and HMan), although the fit appears to be good in both cases. However, for a significant number of variables (namely, aFuc, BasicRegion, MainPeak, G0, Man5, and TotalLMW), the prediction of the BBM is not only significantly worse than that of the CHM, but the model is often unresponsive, as indicated by the almost flat regression line. For such variables, the corresponding relative RMSE is larger than 0.8. Note that, in the case of MainPeak, the regression line has a correct slope due to the presence of a single experimental point and that, thanks to this point only, the relative RMSE is 0.59. In other words, the BBM is often unable to correctly explain the change in the measured CQAs as a function of the manipulated process variables only. This is possibly due to the simplicity of the model, that is, such dependencies can be explained only by taking into consideration of the full (and highly nonlinear) evolution of the process variables.

We then analyzed the robustness of these two models. In Figure S4, in the supporting material, the box plot of the cross-validation values of the BBM (orange) and the CHM (blue) are compared across all Y variables. To perform this analysis, 15 different training sets, each with 36 randomly chosen experiments, that is, with the same size used in analyses in Sections 3.1–3.3, were created, and the remaining experiments were used as test sets. It can be observed that the CHM produces considerably better models than the BBM, which have significantly lower relative RMSE, as indicated by the much smaller value of the box plot median, and much more robust models, as indicated by the narrower distribution of RMSE values for each of the Y

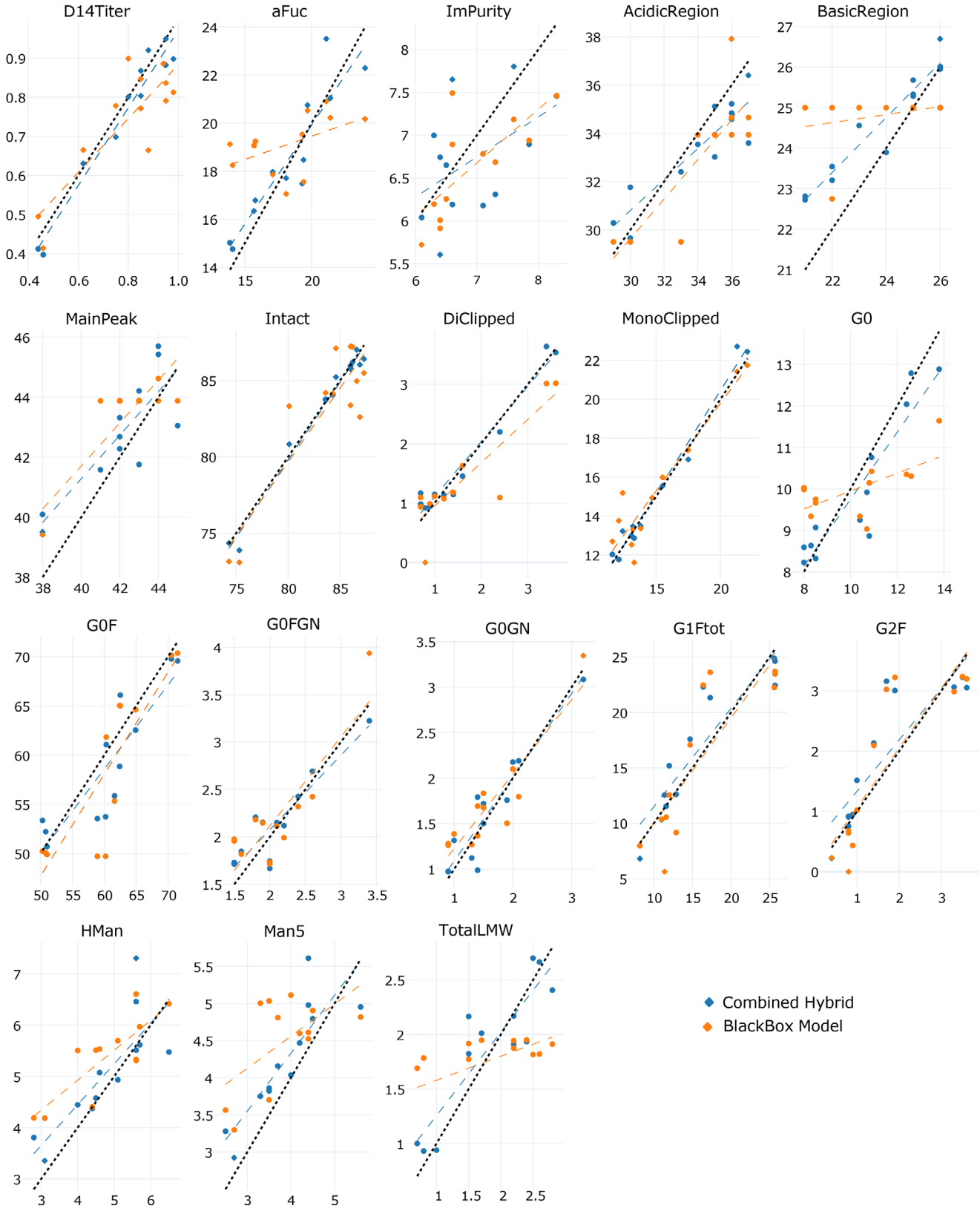

FIGURE 6 Predicted (x-axis) versus experimental (y-axis) value of the different Y variables. Blue dots: combined hybrid model; orange dots: black box model; orange dashed line: "goodness" of the fit line of black box model; blue dashed line: "goodness" of the fit line of the combined hybrid model; and black dotted line: regression fit line.

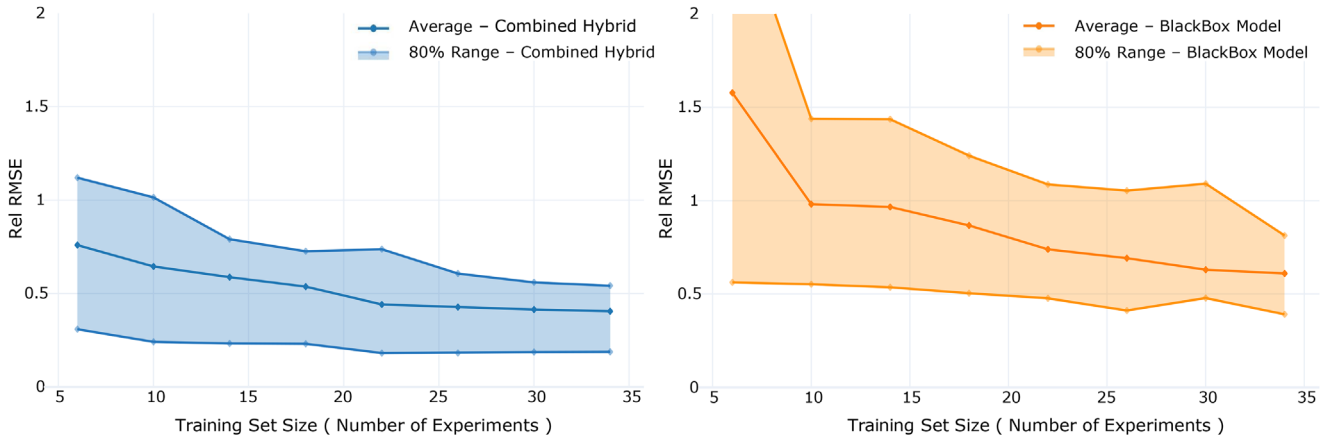

FIGURE 7 Average relative RMSE calculated across all Y variables as a function of the number of experiments in the training set. In all cases, the RMSE was calculated from the same test set, which is used for Figures 2–6. Left figure: combined hybrid model (left); right figure: black-box model. The shadowed area covers 80% of the distribution of relative RMSE values on all Y variables (i.e., from 10 to 90% quantile).

variables. In particular, there are only two Y variables (AcidicRegion and BasicRegion) of the CHM having the median larger than 0.5. In addition, only two variables (AcidicRegion and BasicRegion) have the upper whisker cross the value of 1, which indicates a model without any statistical relevance. Instead, this happens for 13 out of 18 Y variables in the case of the BBM. It should be noted that this analysis does not include the case of the ImPurity, as this variable is not producing reliable HMs (as shown in Figures 7 and 8). In addition to machine learning approaches such as black box models used in this study, mechanistic models are also developed but only for individual product quality like glycans and charge variances.[7,9] These mechanistic models usually require additional process information and complex parameter estimation. In contrast, our model uses only traditional inputs and therefore any product quality attribute as Y variable can be modeled. There is no current literature on similar combined hybrid approaches for simultaneous prediction of full process evolutions linked to product quality.

3.4 | Training efficiency comparison

In this section, we investigated the impact of the number of experiments in the training set on both the precision and robustness of the BBM and CHM. Although linearized models such as the BBM introduce a significant bias in the model, they are known to be robust and less prone to overfitting, especially in the case when very little data is available.[24] For this analysis, using the same algorithm to select the training set as used in previous sections (see Section 2.2.2), we created different training datasets with a reduced number of experiments, from 6 to 34 experiments, while keeping the same test set of 12 experiments as used for Figures 2–6. Note that previous analyses were done with 36 experiments in the training set. Figure 7 shows the calculated average relative RMSE in the test set versus the number of experiments in the training set for the CHM (left, blue) and the BBM (right, orange). In both cases, the average error is computed as in Figure 5, that is, using the relative RMSE on all predicted Y variables. The shaded regions represent the variability of the prediction for all the Y variables: the lower bound represents the 10% quantile of the distribution, while the upper bound represents the 90%. In the case of the CHM, it is noted that, even with only 10 experiments, the model has an average relative RMSE smaller than 0.65, indicating statistical significance on many Y variables, and the relative RMSE for 90% of the Y variables is smaller than 1. The average RMSE drops quickly to reach an almost asymptotic value at 22 experiments (about 0.4, i.e., at the value observed in Figure 5 for 36 experiments). The height of the shaded area, which covers 80% of the variables, is also decreasing and becomes stable when exceeding 30 experiments in the training set, indicating some sensitivity of the CHM model to the size of the training set.

On the other hand, in the case of the BBM, the average of the relative RMSE becomes smaller than 1 only after 18 experiments are included in the training set, and the height of the shaded area is much broader than CHM, indicating less robustness. The average relative RMSE converges to about 0.6, corroborating what is observed in Figure 5, and even when 30 experiments are included in the training set, some variables still have a relative RMSE larger than 1. It is particularly interesting to notice that the final value of the average relative RMSE calculated for the BBM with 34 experiments can be obtained by using 15 experiments only in the training set when using the CHM, thus confirming that the latter is vastly superior in predicting capabilities, in terms of precision and robustness. As discussed in the previous section, this is primarily due to the large number of process variables (15 Z variables) used in this study, which naturally limits the ability of such BBM models to provide reliable predictions when training a small amount of data. In conclusion, the CHM requires only around half of the experiments in model training to match the performance of BBM. However, the required number of experiments depends on factors such as process variation, the number of parameters, their perturbed ranges, and the desired application of the model.

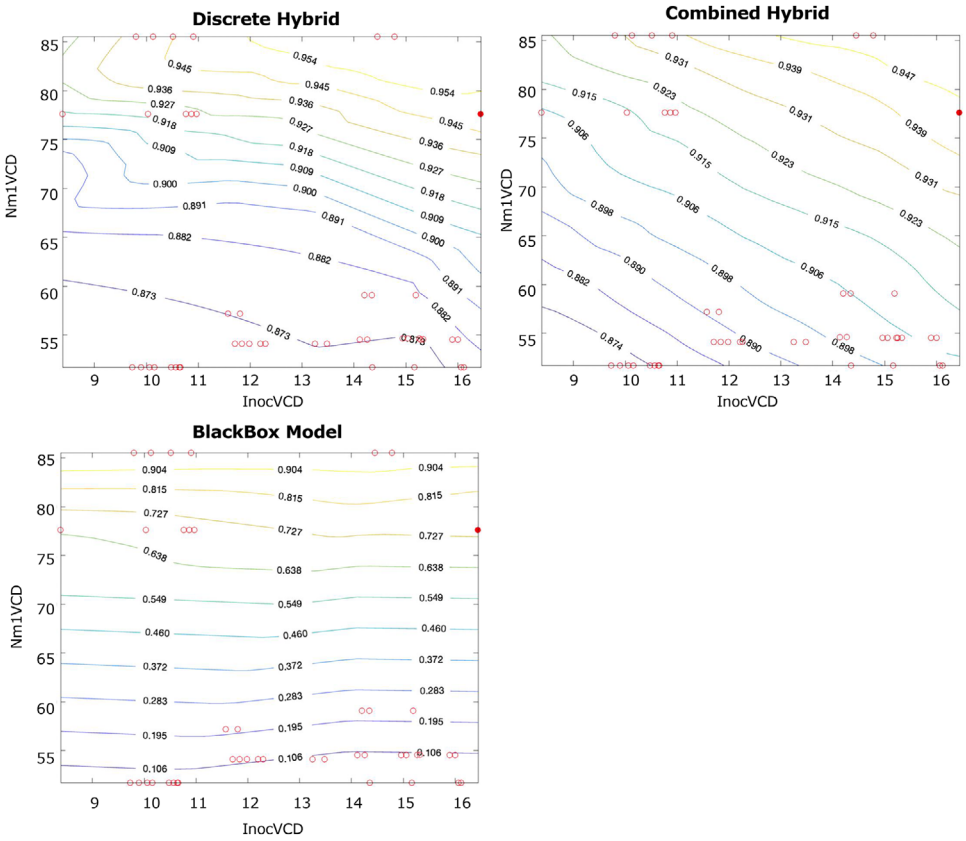

FIGURE 8 Contour plots for normalized final titer for the three studied models as a function of InocVCD and Nm1VCD. PM is in the top left column, CHM is in the top right column, and BBM is in the bottom left column. Empty red circles correspond to all available data and design levels in the dataset. The full red circle corresponds to the condition around which the sensitivity is performed.

3.5 | Process design space characterization

Lastly, we investigated the application of all the models presented in this work for a process characterization study. During process characterization, it is essential to understand how combined changes in the manipulated variables can impact the final product quality in the process design space. In this study, we have restricted such analysis to the evaluation of the final titer. In fact, as seen in Figures 2 and 3, the prediction of the final titer is very precise when using the PM: the relative RMSE on the test set is 0.11, with a replicate relative standard deviation on the final titer of 0.05. Accordingly, we can safely conclude that the PM predicts the "true behavior" of the titer. We decided to compare the prediction of the PM with those obtained by the CHM and the BBM. For this analysis, titer evolution in time was removed from the training of the PM. As shown in Figure 5, the final titer had a relative RMSE in the test set of 0.56 in the BBM, which was significantly larger than the CHM (0.24). Note that this analysis would be practically impossible for any other Y variables due to the limited number of data points, the partial coverage of the design space, and the limited precision of the models. For such variables, we can just observe a disagreement among models, without reasonably being able to indicate which of the models predicts the correct functional behavior even in the case of large differences in the prediction error. The scope of this analysis is to sensibilize how simplified models like the BBM could produce largely distorted predictions in the design space.

Figure 8 shows the contour plots for the normalized final titer for the three models, namely the PM, the CHM, and the BBM, as a function of changes in both Nm1VCD and InocVCD. The solid red dot was used as a reference point, that is, all other variables are left unchanged and equal to this reference experiment. The BBM is almost entirely insensitive to changes to the value of InocVCD, while the PM predicts a slight increase of the final titer when increasing InocVCD. However, the major difference is represented by the effect of Nm1VCD: the BBM produces completely unrealistic values of the final titer, down to a normalized value of about 0.11, while the minimum observed value in the other two models is about 0.87. In the case of the CHM, a more pronounced effect of InocVCD is observed as compared to the PM. However, the span of final titer values with respect to Nm1VCD is very similar in these two cases, indicating a substantial agreement. This simple analysis confirms that the BBM can be highly misleading, thus leading to an incorrect definition of the design space or the need for many experiments to properly characterize the process. The characterization of the process using CHM is not limited to the interaction plots of Figure 8. Other tools to describe the cell culture process using Quality by Design principles[25–27] traditionally used with BBM can also be utilized with CHM.

Furthermore, the required experimentation for CHM is not limited to a traditional DoE, such as full factorial design, and space filling designs, such as Latin Hypercube Sampling, are preferred. As the HM requires the prediction of full process evolution, the accuracy of PM must be sufficiently good in order to use the CHM for CQAs prediction. Though it can be considered that the PM is generally capable of modeling complex process behaviors, aleatoric or epistemic uncertainties might create situations where the model performance does not suffice. In this case, a BBM model could be used as a fallback option. However, the CHM usually benefits from the PM and HM capturing the underlying process state, as the process state is closely linked to the cell state. The changes in the cell state that give rise to changes in the CQAs should hence be captured by the CHM.

4 | CONCLUSION

In this work, we have analyzed the possibility of using hybrid models to estimate the final product quality or CQAs. In previous works, hybrid models, referred to as propagation models here, have been confined to the prediction of the evolution in time of variables like viable cell density or metabolites. This work has again confirmed that robust models can be obtained using hybrid models, with relative RMSE values for the X variables ranging between 0.1 and 0.45 for test sets. However, hybrid models cannot be directly used to predict CQAs, as these are measured only at the end of the process. The CQAs can instead be well predicted by historical models, which integrate the whole evolution of the process variables in time. This study also confirmed that historical models like PLS can predict a broad range of CQAs with relative RMSE values ranging between 0.14 and 0.52 in test sets. In the end, this work demonstrates that a hybrid model can be successfully coupled to a historical model, that is, the combined hybrid model, that directly links process parameters and CQAs. The combined hybrid model outperforms the black-box model by 33% on average in predicting the CQAs. Furthermore, the combined hybrid model requires only around half of the experiments for similar performance when compared to the black box model.

The authors believe that the presented propagation and combined hybrid models can then be used in the future as the basis for creating optimal experimental designs, as these tools can efficiently handle many process variables without exploding the need for experiments, for example, typical factorial designs. In the case where the combined hybrid model's prediction can be trusted (i.e., small prediction interval), the wet-lab experimentation may not be necessary. We can then dedicate the limited experimental resources to only the scenarios where the prediction interval is rather large, meaning the number of wet-lab experiments can be greatly reduced. Moreover, as all capabilities offered by machine learning are embedded, the presented models could be the best candidates to be included in adaptive digital twins, where data can be acquired in real-time, used to continuously refine process knowledge, and deliver predictions for optimal setpoints to control critical process parameters.

AUTHOR CONTRIBUTIONS

Jakub Polak: Conceptualization; formal analysis; investigation; methodology; software; writing – original draft; writing – review and editing; Zhuangrong Huang: Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; software; writing – original draft; writing – review and editing; Michael Sokolov: Conceptualization; formal analysis; investigation; project administration; software; supervision; visualization; Moritz von Stosch: Project administration; supervision; writing – review and editing; Alessandro Butté: Methodology; project administration; supervision; writing – original draft; C. Eric Hodgman: Funding acquisition; writing – review and editing; project administration; resources; supervision; Michael Borys: Writing – review and editing; project administration; resources; supervision; Anurag Khetan: Writing – review and editing; project administration; resources; supervision

ACKNOWLEDGMENTS

All financial support for this collaboration was provided by Bristol Myers Squibb (BMS), USA.

CONFLICT OF INTEREST STATEMENT

JP, MS, MVS, and AB were employees of DataHow AG at the time of study. ZH, EH, MB, and AK were employees of Bristol Myers Squibb at the time of the study.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Zhuangrong Huang https://orcid.org/0000-0002-9145-5582
C. Eric Hodgman https://orcid.org/0009-0001-3837-6084

REFERENCES

1. U.S. Food and Drug Administration. (2004). Pharmaceutical CGMPs for the 21st Century – A Risk Based Approach; Final Report. U.S. Food and Drug Administration: Silver Spring.
2. Grangeia, H. B., Silva, C., Simoes, S. P., & Reis, M. S. (2020). Quality by design in pharmaceutical manufacturing: A systematic review of current status, challenges and future perspectives. *European Journal of Pharmaceutics and Biopharmaceutics*, *147*, 19–37.
3. ICH Q8 (R2). (2009). Pharmaceutical development, ICH Harmonized Tripartite Guidelines. International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use.
4. 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*, 329–336.
5. Simon, L. L., Pataki, H., Marosi, G., Meemken, F., Hungerbuehler, K., Baiker, A., Tummala, S., Glennon, B., Kuentz, M., Steele, G., Kramer, H. J. M., & Rydzak, J. W. (2015). Assessment of recent process analytical technology (PAT) trends: A multiauthor review. *Organic Process Research & Development*, *19*, 3–62.
6. Narayanan, H., Luna, M., Sokolov, M., Butté, A., & Morbidelli, M. (2022b). Hybrid models based on machine learning and an increasing degree of process knowledge: Application to cell culture processes. *Industrial & Engineering Chemistry Research*, *61*, 8658–8672.

7. Karst, D. J., Scibona, E., Serra, E., Bielser, J. M., Souquet, J., Stettler, M., Broly, H., Soos, M., Morbidelli, M., & Villiger, T. K. (2017). Modulation and modeling of monoclonal antibody N-linked glycosylation in mammalian cell perfusion reactors. *Biotechnology and Bioengineering*, *114*, 1978–1990.
8. Rish, A. J., Huang, Z., Siddiquee, K., Xu, J., Anderson, C. A., Borys, M. C., & Khetan, A. (2023a). Identification of cell culture factors influencing afucosylation levels in monoclonal antibodies by partial least-squares regression and variable importance metrics. *Processes*, *11*, 223.
9. Rish, A. J., Siddiquee, K., Huang, Z., Xu, J., Anderson, C. A., Borys, M. C., & Khetan, A. (2023b). Strategies for controlling afucosylation in monoclonal antibodies during upstream manufacturing. *Biotechnology Journal*, *18*, e2200604.
10. Carvalho, M., Riesberg, J., & Budman, H. (2022). Hybrid model to predict the effect of complex media changes in mammalian cell cultures. *Biochemical Engineering Journal*, *186*, 108560.
11. Pinto, J., Mestre, M., Costa, R. S., Striedner, G., & Oliveira, R. (2022). A general deep hybrid model for bioreactor systems: Combining first principles equations with deep neural networks. *Systematic Biology*, *2022*, 495118.
12. Narayanan, H., Behle, L., Luna, M. F., Sokolov, M., Guillén-Gosálbez, G., Morbidelli, M., & Butté, A. (2020). Hybrid-EKF: Hybrid model coupled with extended Kalman filter for real-time monitoring and control of mammalian cell culture. *Biotechnology and Bioengineering*, *117*(9), 2703–2714.
13. Narayanan, H., von Stosch, M., Feidl, F., Sokolov, M., Morbidelli, M., & Butté, A. (2023). Hybrid modeling for biopharmaceutical processes: Advantages, opportunities, and implementation. *Front. Chem. Eng*, *5*, 1157889.
14. Von Stosch, M., Oliveria, R., Peres, J., & De Azevedo, S. F. (2012). Hybrid modeling framework for process analytical technology: Application to Bordetella pertussis cultures. *Biotechnology Progress*, *28*, 284–291.
15. 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*, 773–784.
16. Sokolov, M. (2020). Decision making and risk management in biopharmaceutical engineering—opportunities in the age of covid-19 and digitalization. *Industrial & Engineering Chemistry Research*, *59*, 17587–17592.
17. Rathore, A. S., Pathak, M., Singh, S. K., Read, E. K., Agarabi, C. D., Khan, M., Brorson, K. A., Kumar Singh, S., Pathak, M., Read, E. K., Brorson, K. A., Agarabi, C. D., & Khan, M. (2015). Fermentanomics: Relating quality attributes of a monoclonal antibody to cell culture process variables and raw materials using multivariate data analysis. *Biotechnology Progress*, *31*, 1586–1599.
18. Schmidberger, T., Posch, C., Sasse, A., Gülch, C., & Huber, R. (2015). Progress toward forecasting product quality and quantity of mammalian cell culture processes by performance-based modeling. *Biotechnology Progress*, *31*, 1119–1127.
19. 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*, 1368–1380.
20. 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*, *116*, 2540–2549.
21. Polak, J., von Stosch, M., Sokolov, M., Piccioni, L., Streit, A., Schenkel, B., & Guelat, B. (2023). Hybrid modeling supported development of an industrial small-molecule flow chemistry process. *Computers & Chemical Engineering*, *170*, 108127.
22. Cruz-Bournazou, M. N., Narayanan, H., Fagnani, A., & Butte, A. (2022). Hybrid Gaussian process Models for continuous time series in bolus fed-batch cultures. *IFAC-PapersOnLine*, *55*(7), 204–209.
23. Loeppky, J. L., Moore, L. M., & Williams, B. J. (2010). Batch sequential designs for computer experiments. *Journal of Statistical Planning and Inference*, *140*(6), 1452–1464.
24. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear Regression Analysis*, 6th Edition. John Wiley & Sons, ISBN: 978-1-119-57875-8.
25. Horvath, B., Mun, M., & Laird, M. W. (2010). Characterization of monoclonal antibody cell culture production process using a quality by design approach. *Molecular Biotechnology*, *45*(3), 203–206.
26. Kim, Y. J., Paik, S. H., Han, S. K., Lee, S., Jeong, Y., Kim, J., & Kim, C. W. (2019). Quality by design characterization of the perfusion culture process for recombinant FVIII. *Biologicals*, *59*, 37–46.
27. 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. *Journal of Pharmaceutics and Biopharmaceutics*, *81*(2), 426–437.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Polak, J., Huang, Z., Sokolov, M., von Stosch, M., Butté, A., Hodgman, C. E., Borys, M., & Khetan, A. (2024). An innovative hybrid modeling approach for simultaneous prediction of cell culture process dynamics and product quality. *Biotechnology Journal*, *19*, e2300473. https://doi.org/10.1002/biot.202300473