Bioprocess and Biosystems Engineering (2019) 42:1853–1865
https://doi.org/10.1007/s00449-019-02181-y

RESEARCH PAPER

A bootstrap-aggregated hybrid semi-parametric modeling framework for bioprocess development

José Pinto[1] · Cristiana Rodrigues de Azevedo[1,2] · Rui Oliveira[1] · Moritz von Stosch[1,2,3]

Received: 26 April 2019 / Accepted: 23 July 2019 / Published online: 2 August 2019

Abstract

Hybrid semi-parametric modeling, combining mechanistic and machine-learning methods, has proven to be a powerful method for process development. This paper proposes bootstrap aggregation to increase the predictive power of hybrid semi-parametric models when the process data are obtained by statistical design of experiments. A fed-batch *Escherichia coli* optimization problem is addressed, in which three factors (biomass growth setpoint, temperature, and biomass concentration at induction) were designed statistically to identify optimal cell growth and recombinant protein expression conditions. Synthetic data sets were generated applying three distinct design methods, namely, Box–Behnken, central composite, and Doehlert design. Bootstrap-aggregated hybrid models were developed for the three designs and compared against the respective non-aggregated versions. It is shown that bootstrap aggregation significantly decreases the prediction mean squared error of new batch experiments for all three designs. The number of (best) models to aggregate is a key calibration parameter that needs to be fine-tuned in each problem. The Doehlert design was slightly better than the other designs in the identification of the process optimum. Finally, the availability of several predictions allowed computing error bounds for the different parts of the model, which provides an additional insight into the variation of predictions within the model components.

Keywords Hybrid semi-parametric modeling · Hybrid modeling · Bagging · Design of experiments · Sampling error · Data portioning · Ensemble methods

Introduction

Hybrid semi-parametric models (hereinafter shortly referred to as hybrid models) are a class of models that combine parametric and nonparametric functions in the same mathematical structure [1]. A classic example is the bioreactor dynamic model that combines machine-learning methods such as artificial neural networks (nonparametric) with mass conservation laws (parametric) [1–7]. The mass conservation laws represent well-established scientific knowledge, while machine-learning "learns" unknown (or less understood) cellular kinetics/dynamics from process data. The conceptual advantage is the conjugation of different forms of knowledge/information, which otherwise are not considered together in the same model. The practical advantage is that the number of experiments for process development may be significantly reduced when the underlying model embodies more reliable knowledge at each development step making it more predictive of novel (optimal) process conditions for the next development iteration. Many studies reported different hybrid structures, identification algorithms and process applications reviewed by von Stosch et al. [8]. The most frequently applied machine-learning methods are artificial neural networks (e.g., [2]), partial least squares [9], and more recently support vector machines [10].

The data quantity, quality, and structure are critical for the identification of the nonparametric components of hybrid models. As in standalone nonparametric identification, the data are ideally partitioned into three sets: (1) a training

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00449-019-02181-y) contains supplementary material, which is available to authorized users.

✉ Moritz von Stosch
deq07002@fe.up.pt

[1] REQUIMTE/DQ, Faculty of Science and Technology, University Nova de Lisboa, Campus de Caparica, 2829-516 Caparica, Portugal

[2] CEAM, Faculty of Science, Agriculture and Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, UK

[3] Present Address: Technical R&D, GSK Vaccines, Rue de l'Institut, Rixensart, Belgium

set used to estimate the parameter values; (2) a validation set used to stop parameter estimation and to validate the model; and (3) a test set used to assess the performance of the model on independent data not used for model identification (also sometimes referred to as external validation). For low amounts of data or for data covering distinct process conditions [obtained for instance by Design of Experiments (DoE)], the decision on which data to include in each of the sets is not straightforward. The data-partitioning method can have a significant impact on the model properties and performance, as it can bias/prime the applicability of the model [11]. Heuristics are many times used to guide the data partitioning, namely, (1) the training data should span the explored process conditions; (2) the validation set should contain a repetition experiment to provide a notion for the system noise to variation ratio; and (3) the test data should comprise conditions that evaluate the model performance in the region of interest. However, data partitioning can be addressed more systematically acknowledging the fact that it is essentially a sampling problem, whose effects can be tackled using re-sampling methods. Bagging (short for bootstrap aggregating) is a re-sampling-based ensemble method that was successfully applied to neural networks [12], partial least square models [13, 14], and decision trees [15, 16]. In bagging, the data are re-partitioned several times, one model is developed on every partition, and then, the models are aggregated [17], e.g., by averaging.

In the field of hybrid modeling, bagging and in general ensemble methods have found limited attention to date. Carinhas et al. [14] applied a bagging strategy to obtain the confidence limits for a PLS model in a parallel hybrid structure. Zhang and co-workers showed that the performance of stacked neural networks in hybrid models is superior to standard neural networks [18], stacking being another ensemble method in which the contributions of each neural network to the final prediction are weighted according to their performance on the input domain. A method that simultaneously identifies several parallel nonparametric models and their weighted contributions to the hybrid model has been developed by Peres et al. [19]. This method could be classified as a boosting method, yet another ensemble method. Despite the few reported studies, it seems particularly interesting to adopt and investigate ensemble methods with hybrid models, because they allow evaluating the propagation of uncertainty from one part of the hybrid model into the next, an aspect that has not been studied thus far. This could help to understand the extrapolation performance of hybrid models better, an aspect that is of particular importance for process control and optimization [20].

Methods

Case study: *E. coli* fed-batch process

A previously modeled and optimized *E. coli* fed-batch process serves as case study [21]. A synthetic data set was generated through process simulations, since in this way, the different modeling methods can be impartially compared without the bias of unknown biological and/or experimental variability. Details of the simulation model are provided in "Appendix A". Briefly, the model describes the dynamics of biomass, substrate, and product concentrations in a stirred tank fed-batch bioreactor by applying mass conservation laws. The specific growth, substrate uptake, and product formation kinetic rates are defined as non-linear functions of the substrate concentration and temperature using Monod-type kinetics, where the temperature dependence determines the maximally achievable rates. The design factors were the temperature, T, varying between 29.5 and 33.5 °C, specific growth rate setpoint, μ_{Set}, between 0.1 and 0.16 h^{-1} and biomass concentration at induction, X_{ind}, varying between 5.0 and 19.0 g/kg. This model was simulated for different values of design factors applying three statistical distinct design methods (see Table 1):

Table 1 Different designs of experiments, a short explanation of the DoE, and the respective number of experiments required

Design of experiment	Number of experiments	Number points	Design factors ranges		
			$\mu_{\mathrm{set}}\left(h^{-1}\right)$	T (°C)	$X_{\mathrm{ind}}\left(\frac{\mathrm{g}}{\mathrm{kg}}\right)$
Inscribed central composite design (CCD)	17	249	0.1–0.16	29.5–33.5	5–16.2
Box–Behnken design (BBD)	15	222	0.1–0.16	28.5–33.5	5.0–19.0
Doehlert design (DD)	15	214	0.1–0.16	28.8–33.2	6.3–17.7

The three factors investigated in all DoEs are: biomass concentration at induction ($X_{\mathrm{ind}} = X\left(t_{\mathrm{ind}}\right)$), temperature ($T$) and the desired biomass growth rate (μ_{set}), which was used to compute the exponential feeding rate: $F = 1/S_f \cdot Y_{sX} \cdot \mu_{\mathrm{Set}} \cdot X\left(t_{\mathrm{ind}}\right) \cdot V\left(t_{\mathrm{ind}}\right) \cdot \exp\left(\mu_{\mathrm{Set}} \cdot \left(t - t_{\mathrm{ind}}\right)\right)$. The respective ranges of the factors are: 5–19 (g/l), 29.5–33.5 (C), and 0.1–0.16 (1/h). The center point experiment was repeated three times in each DoE

(1) Inscribed central composite design (CCD) resulted in 17 cultivation experiments (including two repetitions of the center point) and 249 measured points;
(2) Box–Behnken design (BBD) resulted in 15 cultivation experiments and 222 measured points;
(3) Doehlert design (DD) resulted in 15 cultivation experiments and 214 measured points.

The simulated process data were corrupted with 5% white noise. The three synthetic data sets are available as Supplementary file A. These data were used for the hybrid-modeling studies described below.

General bioreactor hybrid model

A stirred tank bioreactor system (as the *E. coli* process described above) can be generically represented by the hybrid model structure, as shown in Fig. 1a. The mass conservation laws are the basis of this model, namely, the dynamic material balance of compounds in an ideally mixed bioreactor:

$$\frac{\mathrm{d}c}{\mathrm{d}t} = K \cdot r(c,x) - D \cdot c + u, \tag{1}$$

with c a vector of concentrations, K a matrix of known yield coefficients, D the dilution rate, u a vector of volumetric feeding rates (control inputs), and $r(c,x)$ a vector of volumetric reaction rates. The latter are complex non-linear functions of the concentrations c and other physicochemical properties x (e.g., temperature and pH). While the structure of mass conservation laws in Eq. (1) is well-established, the reaction rate functions $r(c,x)$ are case dependent and not known with the same level of detail. In the general hybrid model, they are defined as a flexible mixture of parametric and nonparametric functions with the following general form:

$$r(c,x) = r(g(f(c,x),w),c,x). \tag{2}$$

The term $r(g,c,x)$ is a parametric function with well-defined structure based on knowledge. For instance, the Monod cell growth model in case it is applicable. The term $g = g(f(c,x),w)$ is a nonparametric function representing unknown phenomena that needs to be "learned" from data. There are many possibilities to define $g = g(f(c,x),w)$, but the most frequent [8] is a simple feedforward neural network with three layers (also adopted here):

$$g = g(f(c,x),w) = w^{2,1} \cdot \tanh\left(w^{1,1} \cdot f(c,x) + w^{1,2}\right) + w^{2,2}. \tag{3}$$

The transfer functions of the neurons in the input and output layer are linear, while the ones of the hidden layer are hyperbolic tangential. The vector $w = \{w^{1,1}, w^{1,2}, w^{2,1}, w^{2,2}\}$ refers to the parameters that need to be identified from data. The inputs to the nonparametric function typically comprise the concentrations, and/or additional variables, x. For some problems, a non-linear pre-processing function $f(c,x)$ might facilitate the identification of $g(\cdot)$, as, for example, concentration ratios as inputs to a neural network (see, e.g., [22–24]) or other meaningful pre-processing function. The application of this framework to the *E. coli* case study is direct. The main assumption is that the material balance equations are known, while the specific rate equations are

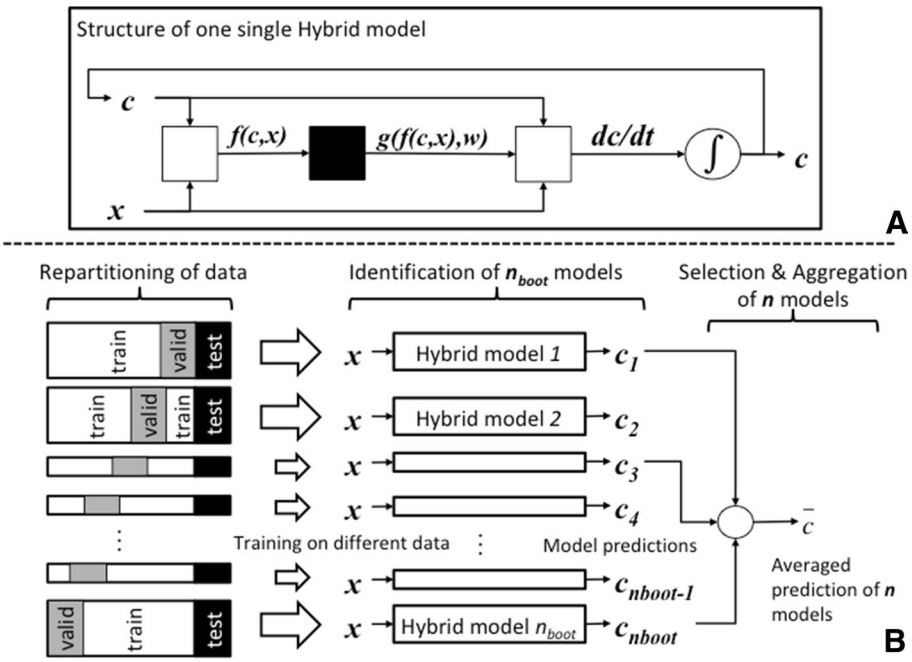

Fig. 1 **a** Hybrid-modeling framework, where the white boxes represent parametric models and the black boxes represent nonparametric models (here Artificial Neural Networks), symbols as in the text. **b** Schematic representation of the bootstrap-aggregated hybrid-modeling structure

unknown. The resulting model equations are provided in "Appendix B". A detailed discussion of the *E. coli* hybrid model is provided elsewhere [21].

Bootstrap-aggregated (bagging) hybrid-modeling framework

The development of hybrid models from small data sets is challenging, because the partitioning of the data may have a significant impact on the performance of the model. Bagging may diminish this impact. It consists in the following three main steps (Fig. 1b):

Step 1: Resample the data contained in the training and validation a given number of times.

Step 2: Parameter estimation resulting in a different model for each pair of training and validation set re-sampled.

Step 3: Aggregate the developed models by averaging their outputs.

Step 1: data re-sampling

The data for hybrid model identification are partitioned into three sets: (1) a training set comprising 2/4 data points used to estimate the parameter values; (2) a validation set comprising 1/4 data points used to stop the parameter estimation and validate the model; and (3) a test set comprising 1/4 of data points used to assess the performance of the model on independent data not used for model identification. The experiments comprised in the test set are kept always the same (more to this in the results section). The experiments in the training-validation sets are randomly re-sampled n_{boot} times from the uniform distribution, yielding n_{boot} training/validation partitions. Re-sampling is performed experimentwise not observationwise. Care must be taken that the exact same validation set is not selected more than once to avoid giving more weight to any particular experimental conditions. It follows that the maximum number of samples is $n_{\text{boot}} = \left(n_{tr} + n_{vd}\right)!/\left(n_{vd}! \cdot n_{tr}!\right)$, with n_{tr} the number of experiments contained in the training set and n_{vd} the number of experiments contained in the validation set.

Step 2: parameter estimation and model validation

A different hybrid model is developed for each of the n_{boot} training/validation samples. The hybrid model structure and size (i.e., the number of hidden nodes in the neural network) is kept always the same. The only variation allowed is in the network parameter values, w. These values are randomly initialized between $[-0.01, 0.01]$ from the Gaussian distribution. Then, parameter estimation is accomplished by minimizing a weighted least squares (WMSE) loss function for the training set (comprising n_{tr} experiments):

$$\min_{w} \left\{ \text{WMSE} = \frac{1}{n_{C \times n_D}} \sum_{i=1}^{n_c} \sum_{j=1}^{n_D} \frac{\left(c_{i,j,\text{mes}} - c_{i,j}\right)^2}{\sigma_i 2} \right\}, \tag{4}$$

where n_c are the number of concentrations, n_D are the number of data points in the training set for concentration, i, $c_{i,j,\text{mes}}$ are the measured concentration values, $c_{i,j}$ are the respective model predictions, and σ_i^2 is the variance computed from the experimental data. The loss function is minimized applying a gradient-based approach, namely, the Levenberg–Marquardt algorithm (Matlab lsqnonlin function). The gradients are obtained using the sensitivities' approach [2, 7]. The WMSE loss function is also monitored for the validation partition comprising n_{vd} experiments. The decision to stop the parameter estimation is made by cross validation, i.e., parameter estimation is stopped when the validation WMSE starts to increase to avoid modeling measurement noise. Since gradient approaches get stuck in local minima, parameter estimation was repeated 100 times from randomly initialized parameter values. From these 100 repetitions, only the best performing model (joint WMSE of the train-validation partition) was selected for aggregation.

Step 3: aggregation

As shown in Fig. 1B, the last stage is the aggregation of the n_{boot} hybrid models by averaging their output variables. In practice, only the $n \ll n_{\text{boot}}$ best hybrid models are aggregated. The models are ranked according to their joint training-validation WMSE. Then, only the n best models are aggregated with n being a design parameter of this framework (discussed in "Results"). The predicted concentrations at a given time instant, t, are then calculated as the mean of concentrations of the n best hybrid models:

$$c_i(t) = \frac{1}{n} \sum_{k=1}^{n} c_i^k(t). \tag{5}$$

The corresponding time-dependent prediction standard deviation, $\sigma_{ci}(t)$, can be computed as

$$\sigma_{ci}(t) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} \left(c_i^k(t) - c_i(t)\right)^2}. \tag{6}$$

Finally, aggregation can be performed at the level of concentrations as given in Eqs. (4) and (5) or at other parts of the model. Since the only part of the model that changes is the nonparametric function $g = g(f(c, x), w)$, aggregation can be performed at this level only, or alternatively at the level of the volumetric reaction rates $r(g, c, x)$. Understanding how the variability propagates from g to r to c is another interesting feature of the bootstrap aggregation framework.

Results and discussion

Process data

Figure 2 shows the variability in the process response originated by the DD, CCD, and BBD. In the case of biomass time profiles (Fig. 2a), the mean concentration of biomass and upper and lower limits ($\pm\sigma$) are remarkably concordant among the DD and CDD. The BDD resulted in a similar average response, but with a much higher dispersion motivated by the fact that the BBD explores more extreme factor values than the other designs particularly in the range of biomass concentration at induction (Table 1). As for the case of substrate concentration time profiles (Fig. 2b), the differences between the designs are more marked, because at the low concentration range explored, the substrate concentration is far more sensitive to the design factors than the other state variables. However, again, the explored region is concordant among the three designs. The more complex abrupt changes in this signal are the result of biomass induction perturbations at different timepoints. Again, the BBD explores a slightly wider region than the other designs. As for the product time profiles (Fig. 2c), the DD and the CDD

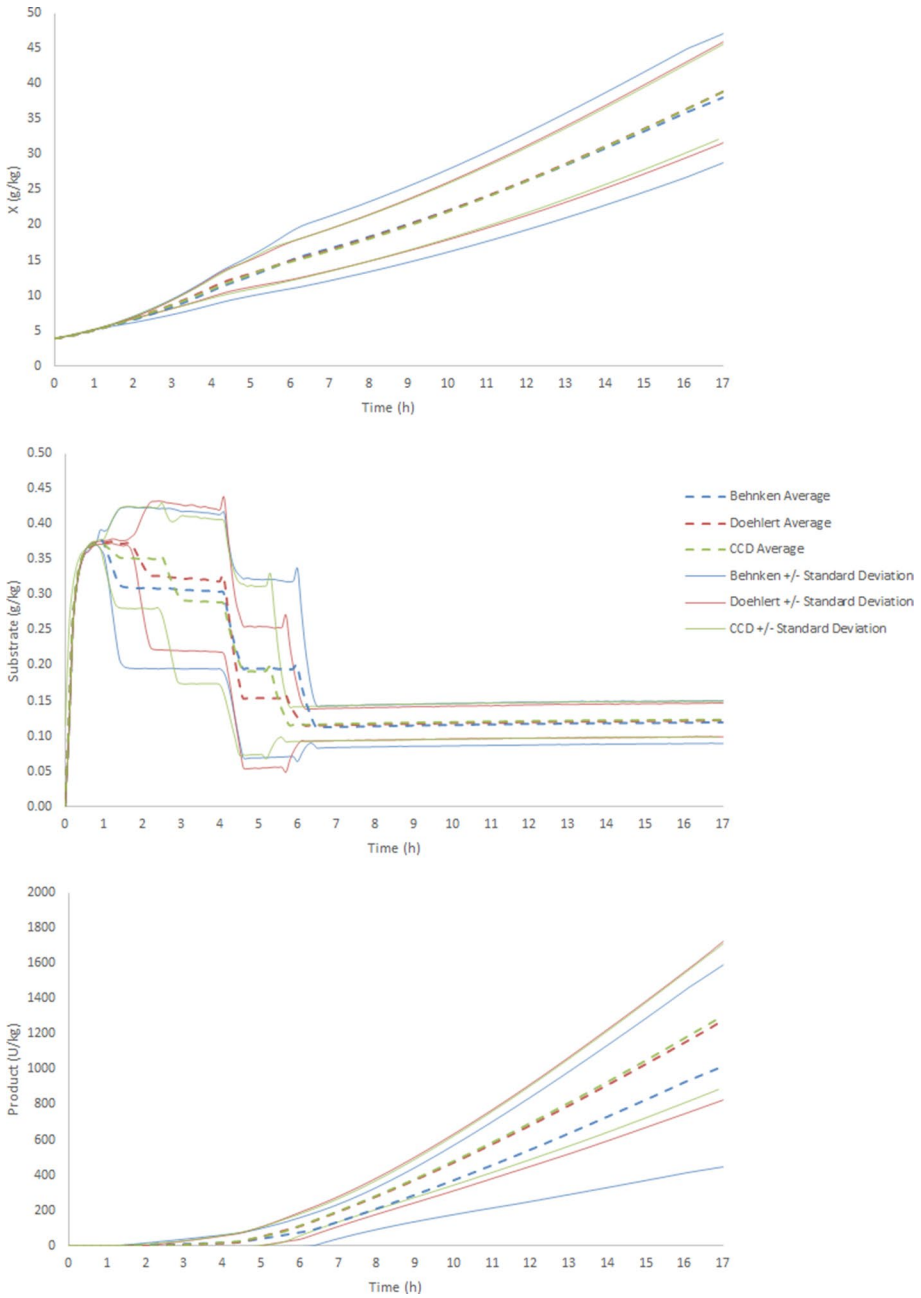

Fig. 2 Synthetic *E. coli* process data generated by model of "Appendix A" and by applying 3 distinct statistical design methods, namely, *CCD* central composite design, *BBD* Box–Behnken design and *DD* Doehlert Design. **a** Biomass concentration over time, **b** substrate concentration over time, **c** product concentration over time. Dashed line: mean profile; full line: ± standard deviation; the different colors represent the different DoE methods. Raw data available as supplementary file A

explore a very similar space as before. The BBD, however, is clearly different in this case, since it explores a much broader space and comparatively lower product concentrations than the other designs. For details in the data, see supplementary file A.

Effect of data partitioning

Hybrid bootstrap aggregation was first applied to the Doehlert data set. Three experiments (13, 14, and 15) were selected for testing, comprising one center point and two extreme experiments with variations in all three design factors (test experiments were always the same). The first 1–12 experiments were selected for training/validation re-sampling with ten experiments for training and two experiments for validation. Re-sampling was performed $n_{boot} = 14$ times resulting in 14 distinct combinations of ten training/two-validation experiments. A hybrid model with the same structure and size was developed for each of the re-sampled data sets using the method previously described. Table 2 summarizes the final modeling error obtained for the $n_{boot} = 14$ hybrid models. The train/valid WMSE varied between 0.0062 (Model 2) and 0.0835 (Model 12), which represents a 13-fold variation. The test WMSE varied between 0.0056 (Model 1) and 0.1630 (Model 12), with a 29-fold variation. The best model was model 2 (lowest combined train/valid/test error), while the worst model was model 12. These results clearly show that data partitioning plays a critical role in the development of the hybrid model. The fact that the data originate from statistical design of experiments intensifies the problem. Statistical design of experiments explores very dissimilar conditions at the lowest number of experiments possible. It is thus very likely that models developed on subsets of such data result to be also very dissimilar.

Effect of aggregation

The hybrid models in Table 2 were ranked from high-to-low train/valid WMSE, and then, only the n top-ranked models (with the lowest train/valid WMSE) were selected for aggregation by averaging the predicted concentrations. Figure 3 shows the effect of n on the train/valid and test WMSE of the aggregated hybrid model. The optimal aggregation number is $n = 4$ in this particular case, which corresponds to the minimum WMSE, as shown in Fig. 3. It becomes clear that aggregating the best $n = 4$ models decreases the train/valid WMSE (16.1% reduction in relation to the best single model), but more importantly, it also decreases the test WMSE by 38% in relation to the best single model. Figure 3 also shows that WMSE increases sharply for $n > 10$. This is the point, where the hybrid models with very large errors are aggregated. They may be seen as "outlier" models that should be removed from the analysis. It is thus very important to drop out the least performing models in the bagging framework. This also makes sense from a sampling point of view, as samples (experiments) that are either based on very different process conditions or that exhibit very different behavior as compared to the other experiments, may keep the training from converging to the overall best performance in cases that these samples are contained in the validations set. This might be the case for data obtained by statistical design of experiments, as the edges of the designs might contain extreme conditions that are not easily extrapolated by the hybrid model.

Table 2 Hybrid-modeling results for the DD data set composed by 15 batch experiments

Model	Train/valid WMSE	Test WMSE	Data partition Train(10)/Valid(2)/[Test](3)
1	0.0078	0.0056	1,2,3,4,5,6,7,8,9,10,11,12,[13],[14],[15]
2*	0.0062	0.0096	1,2,3,4,5,6,7,8,9,10,11,12,[13],[14],[15]
3	0.0086	0.0066	1,2,3,4,5,6,7,8,9,10,11,12,[13],[14],[15]
4	0.0071	0.0092	1,2,3,4,5,6,7,8,9,10,11,12,[13],[14],[15]
5	0.0613	0.1390	1,2,3,4,5,6,7,8,9,10,11,12,[13],[14],[15]
6	0.0084	0.0101	1,2,3,4,5,6,7,8,9,10,11,12,[13],[14],[15]
7	0.0120	0.0113	1,2,3,4,5,6,7,8,9,10,11,12,[13],[14],[15]
8	0.0081	0.0120	1,2,3,4,5,6,7,8,9,10,11,12,[13],[14],[15]
9	0.0126	0.0171	1,2,3,4,5,6,7,8,9,10,11,12,[13],[14],[15]
10	0.0754	0.1326	1,2,3,4,5,6,7,8,9,10,11,12,[13],[14],[15]
11	0.0070	0.0086	1,2,3,4,5,6,7,8,9,10,11,12,[13],[14],[15]
12**	0.0835	0.1630	1,2,3,4,5,6,7,8,9,10,11,12,[13],[14],[15]
13	0.0087	0.0092	1,2,3,4,5,6,7,8,9,10,11,12,[13],[14],[15]
14	0.0617	0.1167	1,2,3,4,5,6,7,8,9,10,11,12,[13],[14],[15]

$n_{boot} = 14$ models were developed by re-sampling ten training batches and two-validation batches. Test batches (13, 14, 15) were always the same

Fig. 3 Weighted mean squared error (WMSE) for the train/valid and test partitions of the Doehlert data set as function of aggregation number, n, i.e., number of best hybrid models to aggregate

Mean and variance of dynamical profiles

One advantage of bagging is the straightforward and automatic calculation of prediction error bounds in a dynamical system. Bagging thus delivers a predicted value (the mean) and also the predicted error bound ($\pm\sigma$ around the mean) along time (Eqs. 4–5). This is here illustrated for the case of DD with $n = 4$ aggregated hybrid models (discussed above). Figure 4 shows the predictions of biomass and product concentrations as well as the respective specific rates for two experiments of the DD-training set. The rates and concentrations' profiles clearly show the switch between the growth and the production phases once the pre-defined induction biomass concentration, X_{ind}, is reached. The error bounds of the specific biomass rates are greater during the growth phase than during the production phase. In addition, while the specific rates vary significantly in some parts, the respective predictions of the concentrations show only minor variations. These observations can partially be explained by the fact that the integration of the material balances has a damping effect on error propagation, which also means that slight differences in the rates do not cause major deviations in the concentrations. However, in the case of the experiment with greater biomass concentration at induction, it can be seen that (1) the error bounds of the concentrations (biomass and product) get wider towards the end of the experiment and (2) the error bounds of the rates are of constant and growing width for specific biomass growth and specific product formation, respectively. This is plausible, as consistent and growing variations in predictions of the reaction rates will ultimately impact on the predictions of the concentrations (accumulating the variation in the predicted concentrations). In case of the product concentrations, it is interesting to note that the multiplication of the specific product formation rate with the predicted biomass concentration could be thought to result into significant variations in the product concentration predictions, which,

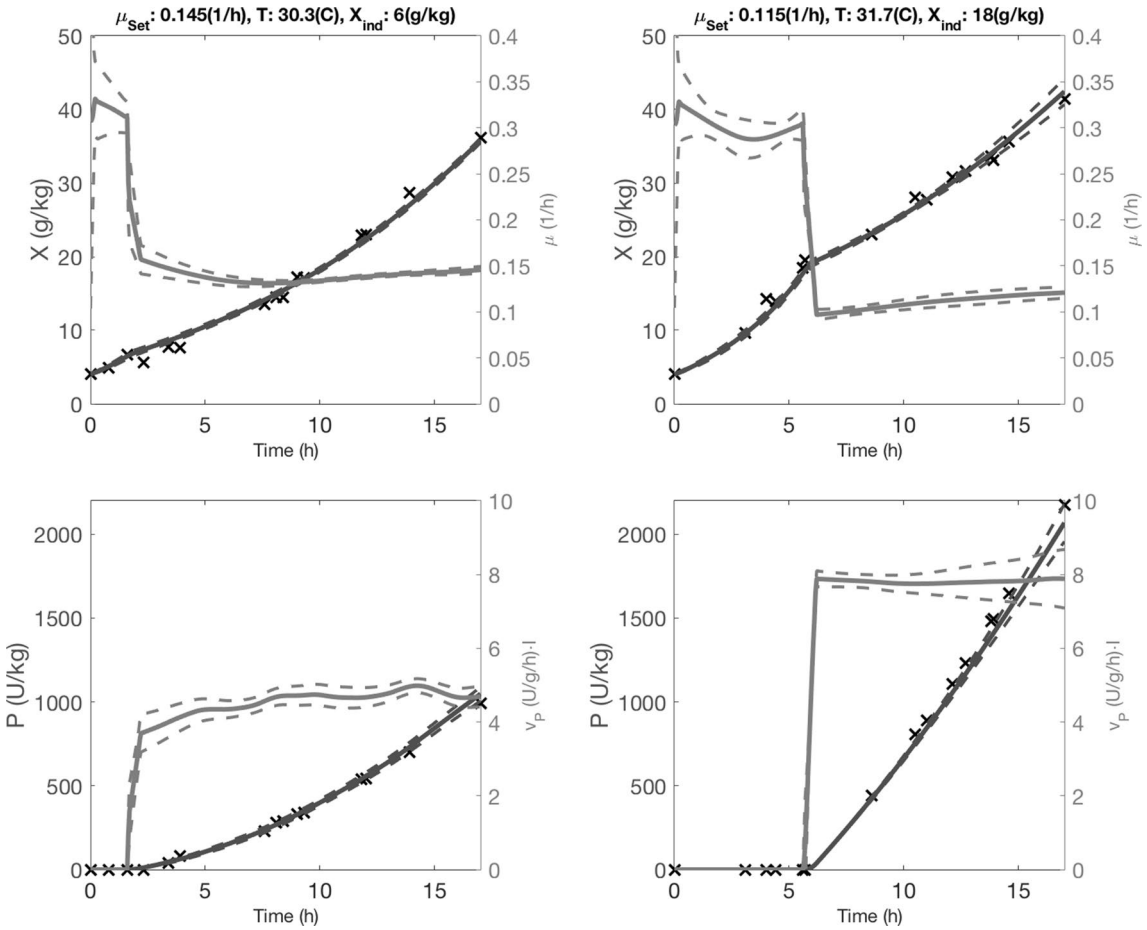

Fig. 4 Experimental and predicted biomass and product concentrations, as well as the predicted specific rates and standard deviations in predicted concentrations and rates over time for two representative experiments from the test set of the Doehlert design

however, is not the case. This might be partially due to the integration.

Is hybrid bagging beneficial?

For a more confident answer to this question, the hybrid bootstrap aggregation framework was extended to the three data sets (DD, CCD, and BBD designs). These designs produce different data sets, thus resulting in different models and different aggregation results. The total number of experiments was 15 for DD and BBD and 17 for the CCD (Table 1). For comparability, ten experiments for training and two experiments for validation were selected in all cases. For the DD and BBD designs, the last three experiments (13–15) were selected for testing, while for the CCD, the last five experiments (13–17) were selected for testing. The number of resampling events was kept the same in all cases ($n_{boot} = 14$), while the optimal aggregation number was investigated in the same way as for the DD design above. The optimal number of best models to aggregate was determined to be four for DD, three for CCD, and three for BBD. Table 3 compares the WMSE of the best single hybrid model with that of the bootstrap-aggregated hybrid model for the three designs. Taking all results together, the bootstrap aggregation methodology improved results across the three designs. Improvements are less expressive in the train/valid partition. Reduction of the WMSE (of aggregated model in relation to the best single mode) was 16.1%, 9.1%, and 2.0% for the DD, BBD, and CCD, respectively. Improvements are particularly significant in the test data set. The reduction of the WMSE (of aggregated model in relation to the best single model) was 38.1%, 51.6%, and 40.0% for the DD, BBD, and CCD, respectively. The substantial reduction of the WMSE in the test partition is particularly meaningful, since it attests the capacity of the final hybrid model to predict new experiments outside the data used for model development.

Table 3 Mean squared error (MSE) for the best hybrid model (BHM) compared with that of the Bootstrap-Aggregated Hybrid Model (BAHM) for the training/validation partition and test partition

DoE	Best model MSE		BAHM MSE	
	Train/valid	Test	Train/valid	Test
CCD	0.0051	0.0210	0.0050	0.0126
BBD	0.0055	0.0632	0.0050	0.0306
DD	0.0062	0.0096	0.0052	0.0059

The models were developed on data from different designs. The optimal numbers of aggregated models were: 5 (CCD), 3 (BBD), and 4 (DD)

CCD inscribed centered composite design, *BBD* Box–Behnken design *DD* Doehlert design

Overall predictive power

Figure 5 shows predicted over measured values of biomass and product for the 3 data sets using the hybrid models of Table 3. Prediction error bounds were calculated and displayed as black bars. The predictions of biomass and product match remarkably well the experimental values in both the training/validation and test partitions for the three data sets (with a few exceptions discussed below). The prediction errors of the test set are slightly higher but comparable to those of the training–validation set.

Comparing the three designs, there are a few differences worth remarking. In the case of the BBD, the magnitude of prediction errors in the test data set is the highest of the three DoEs, which might be explained by the nature of the design. BBD explores extreme experiments that span the space (Fig. 2), but make it difficult to obtain a model that can describe the behavior of the system across the entire space. In this particular case, the aggregation of the models can in principle improve the model performance across the entire space, which was indeed the case with a twofold decrease of WMSE in the test data set. This seems to suggest that the more extreme is the design, the more beneficial might be bootstrap aggregation. In addition, the error bounds are higher in the region of high biomass concentrations (Fig. 4b). These biomass values stem from three different experiments, carried out at very different conditions. It might be that the single hybrid models have not captured the overall behavior of the system due to the changing presence of every one of these experiments in the training and validation set, which lead to the greater error bounds. However, the aggregated model (mean value) seems to capture the behavior of the system well.

In case of the CCD, the model slightly overpredicts one test data experiment exhibiting low product concentrations (Fig. 3a). This experiment was carried out at the lowest design temperature, and therefore, the aggregated model extrapolates, which is known to deteriorate prediction power.

Finally, in case of the DD, the experiment that yielded the highest product concentration is comprised in the test set and the aggregated hybrid model predicts very well the single highest product concentration point (Fig. 3c). This results shows that the model is predictive for product concentrations greater than those observed in the training-validation data, a desired feature when aiming at the maximization of the final product titer.

Identification of process optimum

Methods of design of experiments are routinely applied for process optimization, where the optimum is expected to be found within the process conditions explored by the design (i.e., interpolation of process optimum). Therefore,

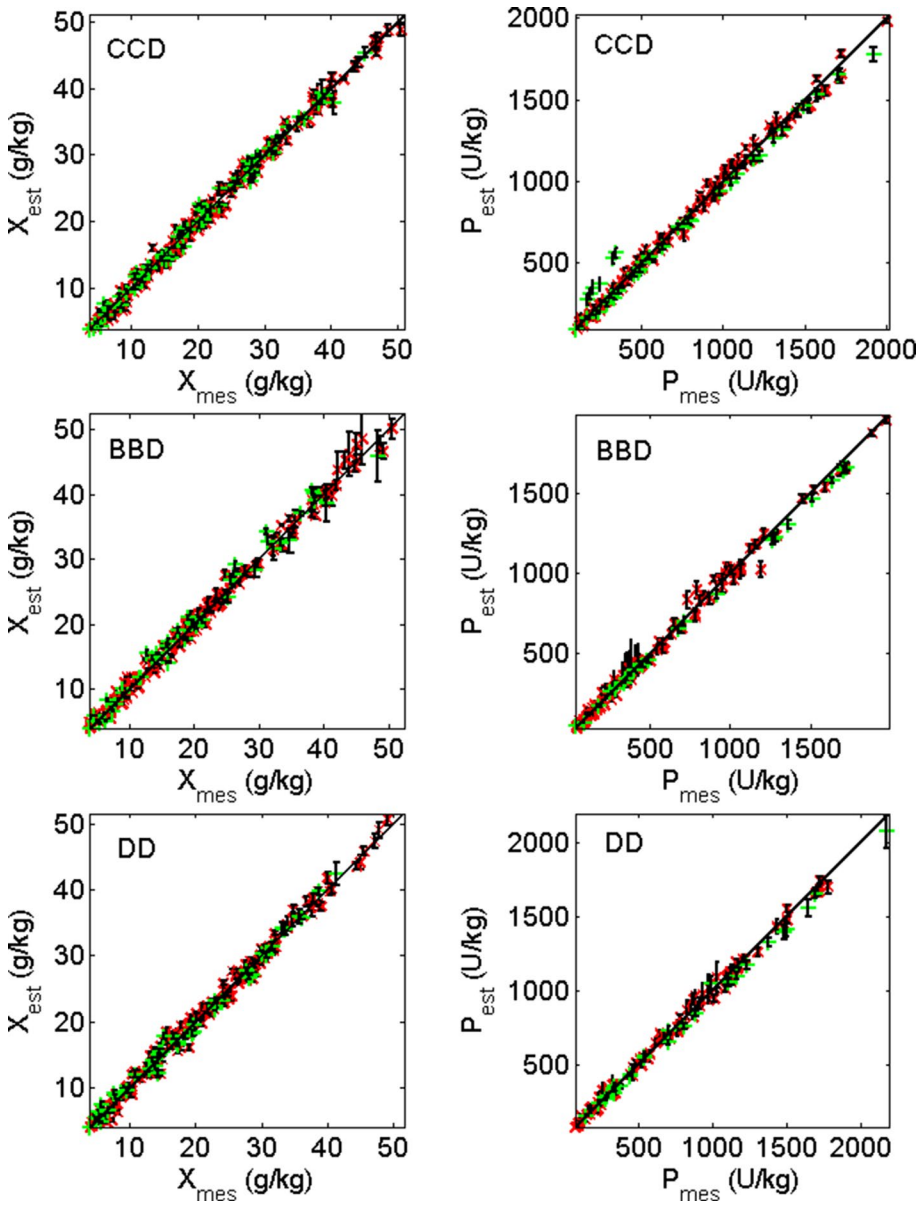

Fig. 5 Predicted biomass and product concentrations over experimental measured concentrations for the bootstrap-aggregated hybrid models developed on data originating from the *CCD* central composite design, *BBD* Box–Behnken design, and *DD* Doehlert design. Red x: training and validation set; green +: test set; black bars: standard deviations of the three predictions

it is interesting to compare the final product concentration predictions of the three aggregated hybrid models of Table 3. Figure 6 shows the "true" final product concentrations (at cultivation time of 17 h) as a function of explored design space (Fig. 6a) and the respective predictions by the bootstrap-aggregated hybrid models derived from CCD, BBD, and DD data sets (Fig. 6b–d). It can be observed that all three aggregated hybrid models correctly indicate the process region in which the highest product concentrations can be found. However, the shape of this region and the accuracy of the predicted concentrations are only conserved well for the CCD- and DD-bootstrap-aggregated hybrid models (BAHM). The CCD-BAHM describes the product concentrations within the overall space best, which could be expected, as the space had been well characterized by the experiments, as described above. The DD-BAHM seems to describe the behavior of the process towards the limits of the investigated ranges better than the other designs and the predictions across the overall space are good. It seems that the space-filling manner in which the DD explores the space helps the hybrid model to learn the systems behavior, which is also in agreement with the findings of other researchers for other modeling techniques [25, 26].

Sometimes, the process optimum may lay outside the explored design space. In the problem studied here, the conditions that maximize the final product titer are outside of the explored process region, namely, at a higher biomass induction concentration (25 g/kg at optimum, studied range 5–19 g/kg). This was done on purpose to assess the extrapolation capabilities of the hybrid-modeling framework.

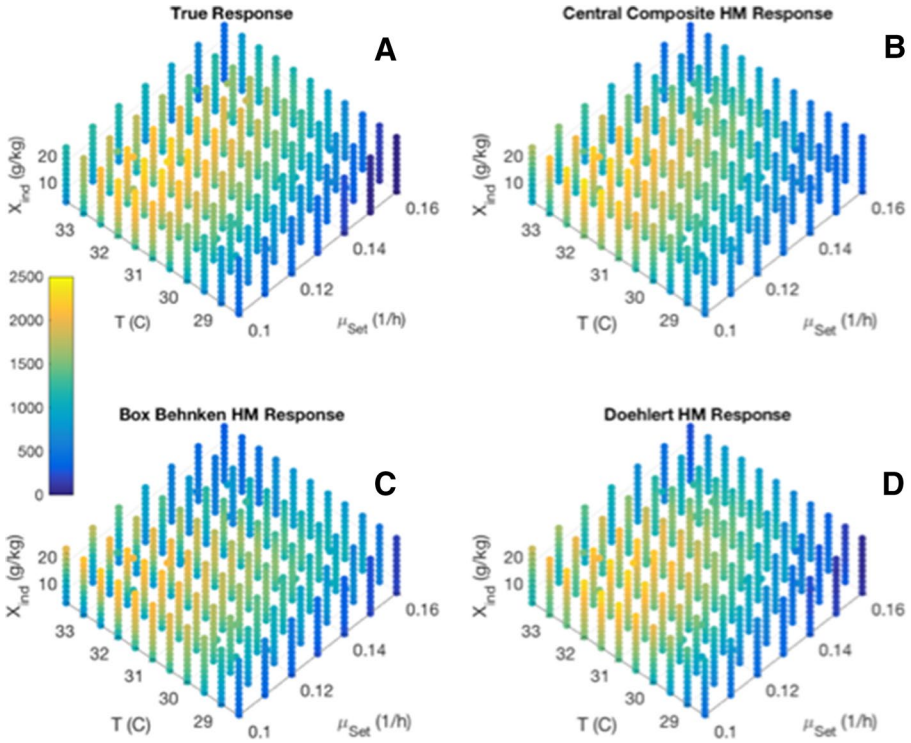

Fig. 6 Final time product concentrations ($t_f = 17$ h) over process conditions for **a** true process model; **b** CCD-; **c** BBD-, and **d** DD bootstrap-aggregated hybrid models. Note that the process conditions reach into a process region, i.e., initial biomass concentrations greater than 19 g/kg, that were not covered in the original DoEs and, therefore, has explorative character

Studying the impact of these changes seems particularly interesting, as on one hand miss predictions of biomass concentration cause subsequent miss predictions in all other compounds (it is multiplied with the specific rates, wherefore it impacts on the evolution of all concentrations), but on the other hand, it also is an input to the nonparametric model (which typically does not extrapolate well). Figure 7 shows the predictions of the three aggregated hybrid models and the true response surface for the final product concentrations at the optimal biomass induction concentration (25 g/kg). It can be seen that the predicted response surfaces of all three models agree fairly well with the true response surface. Apart from the BBD-BAHM, which suggest that the optimal conditions are outside the investigated ones (for temperature and feeding rate), the optimal conditions are fairly accurately captured by the models. The error bounds provide an indication of the reliability of the predictions and it can be seen that the CCD-BAHM provides the most reliable predictions followed by the DD-BAHM.

Conclusions

This study investigated a hybrid-modeling method that combines mechanistic modeling with machine-learning methods to extract knowledge from data. In particular, a bootstrap-aggregated hybrid-modeling framework was studied to reduce the bias of the training and validation data choice, which is particular pronounced for small data sets with distinct changes in conditions, such as data obtained from statistical design of experiments. Three synthetic data sets of an *E. coli* fed-batch process generated by three distinct designs (central composite design, Box–Behnken design, and Doehlert design) were used to compare the different methods. Taking all results together, it may be concluded that the proposed bootstrap aggregation framework significantly increases the predictive power of hybrid semi-parametric models when the data are obtained by statistical design of experiments. This advantage is vital in a context of bioprocess development, because optimal operating conditions can be more accurately predicted by the hybrid model at each process development stage, thus globally reducing the experimental effort for process development. The ability to easily compute reliable error bounds of a dynamical system and for different model parts is particularly interesting for process monitoring and process optimization/control. Online decisions can be made not only on the basis of a profit function, but also on a quantitative measure of the reliability of predictions. The main downside is of course the increased computation time, which can be roughly estimated to be n_{boot} times higher (in our case 14 times higher) when compared to not doing bootstrap aggregation. However, with the ever-increasing computation power, this disadvantage is not seen as a severe limitation in practice. In the future, it should

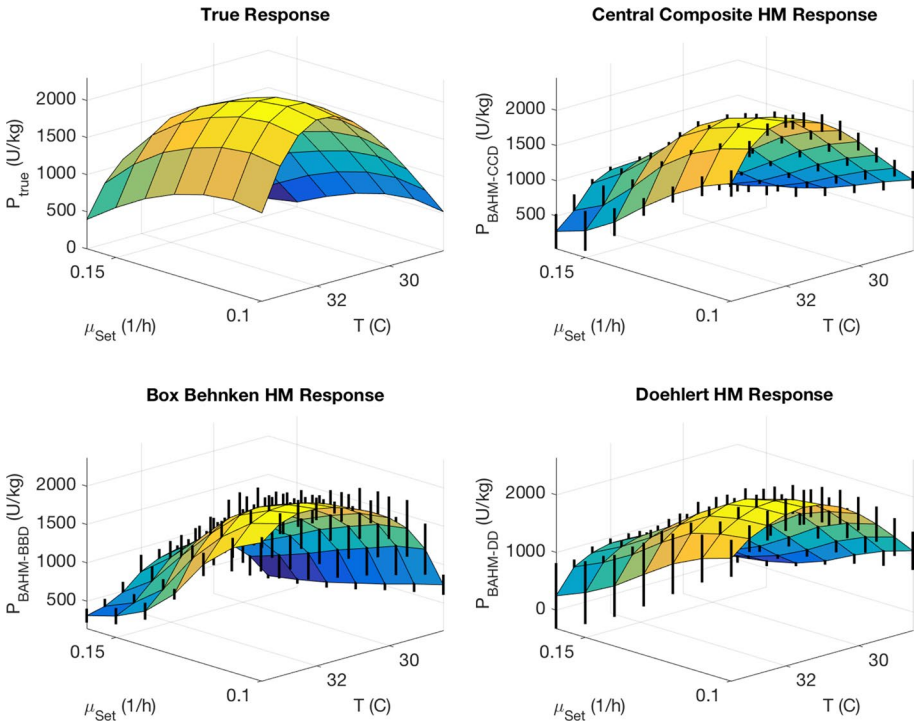

Fig. 7 Response surface of the final time product concentrations ($t_f = 17$ h) at an initial biomass concentration of 25 g/kg obtained from the true process model and the CCD-, BBD-, and DD bootstrap-aggregated hybrid models. In addition, the standard deviations of the predictions are presented

be studied whether other aggregation methods than averaging, e.g., data-domain-specific model prediction weighting (stacking or boosting alike) can improve the prediction performance even further.

Compliance with ethical standards

Conflict of interest The other authors declare that no competing interests exist.

Appendix

A: *E. coli* simulation fed-batch model

The model describes the production of viral capsid protein by a recombinant *E. coli* strain in a fed-batch bioreactor. This model has been proposed by [21], which is an adaptation of the model by [27]. The model comprises the material balances for biomass, substrate, and product concentration as well as the overall mass balance in a stirred tank bioreactor:

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \mu \cdot X - D \cdot X, \tag{7}$$

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -v_S \cdot X - D \cdot \left(S - S_f\right), \tag{8}$$

$$\frac{\mathrm{d}P}{\mathrm{d}t} = v_P \cdot X - D \cdot P, \tag{9}$$

$$\frac{\mathrm{d}W}{\mathrm{d}t} = u_F, \tag{10}$$

with μ, v_S, and v_P the specific rates of biomass growth (1/h), substrate uptake (1/h), and product formation (U/g/h), X, S, and P the biomass (g/kg), substrate (g/kg), and product concentrations (U/kg), $D = u_F/W$ (1/h) the dilution rate, and u_F the feeding rate (kg/h).

The specific biomass growth rate was modeled using the expression:

$$\mu = \mu_{\max} \cdot \frac{S}{S + K_S} \cdot \frac{K_i}{S + K_i} \cdot \exp\left(\alpha \cdot \left(T - T_{\mathrm{ref}}\right)\right), \tag{11}$$

with $\mu_{\max} = 0.737$ (1/h), $K_S = 0.00333$ (g/kg), $K_i = 93.8$ (g/kg), $\alpha = 0.0495$ (1/C), $T_{\mathrm{ref}} = 37$ (°C), and T (°C) the temperature of the culture broth.

The specific substrate uptake rate is modeled via:

$$v_S = \frac{1}{Y_{XS}} \cdot \mu + m, \tag{12}$$

with $Y_{XS} = 0.46$ (g/g) and $m = 0.0242$ (g/g/h).

The specific product formation rate is modeled by

$$v_P = \frac{I_D}{T_{\mathrm{PX}}} \cdot \left(\frac{v_{P,\max,T} \cdot \mu \cdot k_m}{k_\mu + \mu + \mu^2/k_{i\mu}} - p_X \right), \tag{13}$$

with

$$v_{P,\max,T} = \frac{5 \cdot 10^{10} \cdot \exp\left(\frac{-A_{\mathrm{eng}}}{R \cdot (T + 273.15)}\right)}{1 + 3 \cdot 10^{93} \cdot \exp\left(\frac{-R_{\mathrm{eng}}}{R \cdot (T + 273.15)}\right)}, \tag{14}$$

with $A_{\text{eng}} = 62$ (kJ/mol), $R_{\text{eng}} = 551$ (kJ/mol), $R = 8.3144e-3$ (kJ/mol/K), $T_{PX} = 1.495$(h), $p_X = 50$ (U/g), $k_\mu = 0.61$(1/h), $k_m = 751$(U/g), $k_{i\mu} = 0.0174$ (1/h), and the induction parameter $I_D = 0$ before induction and $I_D = 1$ afterwards.

For the feeding rate, an exponential profile was adopted to match a desired constant specific biomass growth, μ_{set}, that is

$$u_F = \frac{1}{S_f \cdot Y_{XS}} \cdot \mu_{\text{set}} \cdot X_0 \cdot W_0 \cdot \exp\left(\mu_{\text{set}} \cdot \left(t - t_0\right)\right), \tag{15}$$

where $X_0 = X\left(t_0\right)$ (g/kg) is the initial biomass concentration and $W_0 = W\left(t_0\right)$ (kg) is the initial weight of the culture broth.

The process was divided into two phases, a growth and a production phase. During the growth phase, $\mu_{\text{set}} = 0.3$ (h^{-1}) and $T = 34$ (C). The duration of the growth phase was adapted to yield the initial biomass concentration, X_{ind}, set out by the DoEs. The substrate concentration in the feeding solution was set to $S_f = 300$ (g/kg). Data for online variables were logged every 6 min. The biomass and product concentrations (offline variables) were measured 20 times during each fermentation. The data were corrupted with 5% Gaussian (white) noise.

B: *E. coli* hybrid semi-parametric model

The parametric part of the hybrid model is based on the material balance equations of biomass and product, that is

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \mu \cdot X - D \cdot X, \tag{16}$$

$$\frac{\mathrm{d}P}{\mathrm{d}t} = v_p \cdot X - D \cdot P, \tag{17}$$

where D is the dilution rate, X and P are the biomass and product concentrations (to note that the hybrid model does not consider substrate dynamics), with specific reaction rates μ and v_p. Thus, the volumetric rate Eq. (2) simplifies as follows for the present problem:

$$r(c,x) = \left[\mu, v_p\right]^{\mathrm{T}} \cdot X. \tag{18}$$

The specific rates μ and v_p are much more difficult to establish; thus, they were modeled by a simple feedforward neural network with three layers only:

$$g = \left[\mu, v_p\right]^{\mathrm{T}} = w^{2,1} \cdot \tanh\left(w^{1,1} \cdot f(c,x) + w^{1,2}\right) + w^{2,2}, \tag{19}$$

with $w = \{w^{1,1}, w^{1,2}, w^{2,2}, w^{2,2}\}$. The network has only three inputs, namely, biomass, X, the feeding rate, F, and cultivation temperature, T. Thus, the pre-processing function (Eq. (3)) reduces to the following form:

$$f(c,x) = [X, F, T]^{\mathrm{T}}. \tag{20}$$

Preliminary tests have shown that five neurons in the hidden layer are optimal for the present case study data set used, which corresponds to $\dim(w) = 4 \times 5 + 6 \times 2 = 32$ parameters to be identified in each run. The number of hidden nodes of the neural network was thus selected to be five in all studies performed.

References

1. Thompson ML, Kramer MA (1994) Modeling chemical processes using prior knowledge and neural networks. AIChE J 40:1328–1340. https://doi.org/10.1002/aic.690400806
2. Psichogios DC, Ungar LH (1992) A hybrid neural network-first principles approach to process modeling. AIChE J 38:1499–1511. https://doi.org/10.1002/aic.690381003
3. Simutis R, Oliveira R, Manikowski M et al (1997) How to increase the performance of models for process optimization and control. J Biotechnol 59:73–89. https://doi.org/10.1016/S0168-1656(97)00166-1
4. Schubert J, Simutis R, Dors M et al (1994) Bioprocess optimization and control: application of hybrid modelling. J Biotechnol 35:51–68. https://doi.org/10.1016/0168-1656(94)90189-9
5. van Can HJL, te Braake HAB, Bijman A et al (1999) An efficient model development strategy for bioprocesses based on neural networks in macroscopic balances: part II. Biotechnol Bioeng 62:666–680. https://doi.org/10.1002/(SICI)1097-0290(19990320)62:6%3c666:AID-BIT6%3e3.0.CO;2-S
6. von Stosch M, Peres J, de Azevedo SF, Oliveira R (2010) Modelling biochemical networks with intrinsic time delays: a hybrid semi-parametric approach. BMC Syst Biol 4:131. https://doi.org/10.1186/1752-0509-4-131
7. Oliveira R (2003) Combining first principles modelling and artificial neural networks: a general framework. Comput Aided Chem Eng 14:821–826. https://doi.org/10.1016/S1570-7946(03)80218-3
8. von Stosch M, Oliveira R, Peres J, Feyo de Azevedo S (2014) Hybrid semi-parametric modeling in process systems engineering: past, present and future. Comput Chem Eng 60:86–101. https://doi.org/10.1016/J.COMPCHEMENG.2013.08.008
9. 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 Syst Appl 38:10862–10874. https://doi.org/10.1016/J.ESWA.2011.02.117
10. Wang X, Chen J, Liu C, Pan F (2010) Hybrid modeling of penicillin fermentation process based on least square support vector machine. Chem Eng Res Des 88:415–420. https://doi.org/10.1016/J.CHERD.2009.08.010
11. Portela RMC, von Stosch M, Oliveira R (2018) Hybrid semiparametric systems for quantitative sequence-activity modeling of synthetic biological parts. Synth Biol 3:10
12. Zhang J (1999) Developing robust non-linear models through bootstrap aggregated neural networks. Neurocomputing 25:93–113. https://doi.org/10.1016/S0925-2312(99)00054-5
13. Mevik B-H, Segtnan VH, Næs T (2004) Ensemble methods and partial least squares regression. J Chemom 18:498–507. https://doi.org/10.1002/cem.895
14. Carinhas N, Bernal V, Teixeira AP et al (2011) Hybrid metabolic flux analysis: combining stoichiometric and statistical constraints to model the formation of complex recombinant products. BMC Syst Biol 5:34. https://doi.org/10.1186/1752-0509-5-34

15. Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9:181–199. https://doi.org/10.1007/s10021-005-0054-1
16. Svetnik V, Liaw A, Tong C et al (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 43:1947–1958. https://doi.org/10.1021/ci034160g
17. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140. https://doi.org/10.1007/BF00058655
18. Tian Y, Zhang J, Morris J (2004) Dynamic on-line reoptimization control of a batch MMA polymerization reactor using hybrid neural network models. Chem Eng Technol 27:1030–1038. https://doi.org/10.1002/ceat.200402068
19. Peres J, Oliveira R, Feyo de Azevedo S (2000) Knowledge based modular networks for process modelling and control. Comput Aided Chem Eng 8:247–252. https://doi.org/10.1016/S1570-7946(00)80043-7
20. Kahrs O, Marquardt W (2007) The validity domain of hybrid models and its application in process optimization. Chem Eng Process Process Intensif 46:1054–1066. https://doi.org/10.1016/j.cep.2007.02.031
21. von Stosch M, Willis MJ (2017) Intensified design of experiments for upstream bioreactors. Eng Life Sci 17:1173–1184. https://doi.org/10.1002/elsc.201600037
22. 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 Biosyst Eng 39:773–784. https://doi.org/10.1007/s00449-016-1557-1
23. Gnoth S, Simutis R, Lübbert A (2010) Selective expression of the soluble product fraction in *Escherichia coli* cultures employed in recombinant protein production processes. Appl Microbiol Biotechnol 87:2047–2058. https://doi.org/10.1007/s00253-010-2608-1
24. Gnoth S, Jenzsch M, Simutis R, Lübbert A (2008) Product formation kinetics in genetically modified *E. coli* bacteria: inclusion body formation. Bioprocess Biosyst Eng 31:41–46. https://doi.org/10.1007/s00449-007-0161-9
25. Lin Y, Zhang Z, Thibault J (2009) Comparison of experimental designs using neural networks. Can J Chem Eng 87:965–971. https://doi.org/10.1002/cjce.20233
26. Alam FM, McNaught KR, Ringrose TJ (2004) A comparison of experimental designs in the development of a neural network simulation metamodel. In: Simulation modelling practice and theory. pp 559–578
27. Levisauskas D, Galvanauskas V, Henrich S et al (2003) Model-based optimization of viral capsid protein production in fed-batch culture of recombinant *Escherichia coli*. Bioprocess Biosyst Eng 25:255–262. https://doi.org/10.1007/s00449-002-0305-x