Hybrid Modeling and Intensified DoE: An Approach to Accelerate Upstream Process Characterization

Benjamin Bayer, Gerald Striedner, and Mark Duerkop*
variables, the so-called principal components (PC), in the investigated spectra can be determined. Generally, there are two different modeling approaches: nonparametric (black-box) and parametric (white-box) models. Nonparametric models are built on experimental data only and do not need any further process knowledge. Various regression techniques are available and commonly applied to develop nonparametric models. In contrast, parametric models use empirical knowledge and first principles, that is, their structure is well defined and transparent. Both modeling approaches possess separate unique advantages as well as disadvantages and limitations due to their respective model structures.

1.3. Hybrid Modeling

The concept of combining a nonparametric and a parametric model into a single semi-parametric model structure is called hybrid modeling. This allows the incorporation of both process knowledge and data-driven information. The hybrid model structure overcomes the shortcomings of each separate modeling technique, for example, the black box can be used to calculate parameters in the white box, which therefore do not have to be solely assumed, reducing errors at the cost of increased complexity. For instance, the values of specific rate expressions are known unknowns a priori to a bioprocess and must first be determined, for example, by process modeling. However, by solely utilizing a white-box model, these rate values must be assumed from data using a defined causal method, but in a hybrid model, they first can be estimated in a defined black box and then transferred to the white box.

For example, by utilizing process variables that have an influence on these rates as input to the black box, the incorporation of this impact can also be taken into account in the white box, generating hybrid model predictions closer to the analytical values. Artificial neural networks (ANN) are frequently utilized for this process. Accurate rate estimations are of great importance for robust bioprocess modeling, and achieving these as precisely as possible is of high interest. Due to these advantages, hybrid modeling is gaining in popularity for bioprocess modeling. Even though a hybrid model provides improved performance compared to other approaches, the possibility of misprediction still exists. To ascertain the chance of such model uncertainty, cross validation is commonly performed in machine learning to calculate the average misprediction possibility. However, bootstrapping can also be applied for this task and has been proven to be a more flexible technique, allowing full control over developing the final model. In this method, a number of models are merged into one, which leads to a probability of misinterpretation and risk assessment occurring from different data permutations. However, these techniques are linked to an increased computational workload, since several hybrid models must be developed. This workload linearly increases, either with the number of chosen folds for the cross validation or the number of applied bootstraps.

The combination of both elements, hybrid modeling and bootstrapping, provides a robust and reliable hybrid model for bioprocess modeling. Nevertheless, the experimental workload, that is, the generation of the required process data and the related analytical effort, rapidly ends up being laborious and time consuming.

1.4. Intensified Design of Experiments

A promising approach to reducing the experimental workload is to change the CPPs during the cultivation. With these intra-experimental CPP set-point changes, the reaction to dynamic changes in the process can be captured. By performing this intensified DoE (iDoE), one should note that the history of the cell contributing to a memory effect is often assumed to influence how the cells react to subsequent CPP set points. One of the main challenges is to describe the process dynamics in response to intra-experimental changes and to estimate the behavior of the cells under constant conditions. Therefore, a time-resolved hybrid model can be built on iDoE data to describe the occurring process dynamics, because it captures the whole process. This emphasizes a combinatorial approach, using hybrid modeling and iDoE, to generate process knowledge and simultaneously accelerate process characterization.

1.5. A Combinatorial Approach Leading to Accelerated Process Development

To significantly reduce the number of required experiments for developing a process model, we present the concept of iDoE. As basis for comparison, we used a completely characterized three-dimensional design space of a previously derived Escherichia coli (E. coli) fed-batch study at the 20L scale with 27 distinct CPP combination settings. The intensified experiments were performed in the same design space but contained two CPP set-point shifts during each cultivation, so that three CPP combination settings were tested overall within one fed-batch fermentation. This led to nine iDoE cultivations to completely characterize the same design space. Consecutively, to examine a possible memory effect of the cells, the online process data and the 2D-fluorescence spectra of the static and intensified fed-batch fermentations were investigated and compared using exploratory data analysis (PCA and PARAFAC) to test for any differences.

A hybrid model was built on iDoE data, and its performance was compared to a previously developed full-factorial static hybrid model, built on the complete design space. To further challenge the iDoE approach, a fractional-factorial static hybrid model, built only on the center point and corners of the design space, was assessed to challenge the potential time reduction and advantages regarding process characterization using iDoE.

2. Experimental Section

2.1. Experimental Design

For all fed-batch cultivations, E. coli (HMS174 (DE3)) was utilized for expressing recombinant human superoxide dismutase at the 20 L scale. The experimental design consisted of a full-factorial design space with three CPPs: the specific growth rate (µ) controlled by the substrate feeding rate, the cultivation temperature
respectively, are \( \mu = 0.10, 0.15, \) and 0.20 h\(^{-1} \); \( T = 30, 34, \) and 37 °C; and \( I = 0.2, 0.5, \) and 0.9 \( \mu \)mol IPTG g\(^{-1} \) cell dry mass. This results in 27 CPP combination settings, as presented elsewhere. \(^{[26]} \) The complete list of all performed fed-batch cultivations of the DoE is given in Table S1, Supporting Information. For all these cultivations in the design space, the analytical measurements for the biomass concentration (in g L\(^{-1} \)) and the soluble product titer (in g L\(^{-1} \)) were assessed by thermogravimetric analysis\(^{[27]} \) and ELISA\(^{[28]} \) respectively. The analytical error of the biomass and product titer determination was assessed from seven replicate runs in a previous study\(^{[20]} \) with 3.6% and 7.6%, respectively. 

The fed-batch phase was always carried out for four doubling times. The induction of the cells took place after the first doubling time, enabling recombinant protein production for three doubling times. All information about the applied exponential feeding strategy for the fed-batch phase, the utilized \( E. \) coli strain, the expression vector system, the online monitoring, and the offline measurements were presented elsewhere.\(^{[25–31]} \) In addition to the standard online available process variables, such as pH, temperature, inlet air, stirrer speed, base consumption, accumulated feed, inducer, and head pressure, a 2D fluorescence probe (BioView, Delta Light and Optics, Denmark) was utilized to measure the cultivation broth in 20 nm steps (from ex270/em310 up to ex550/em590), resulting in 120 excitation/emission wavelength variables. These measurements were used to examine if differences on the cellular and process level are visible between the DoE and iDoE.

To investigate and quantify the metabolic burden, possible toxicity, and induced stress due to recombinant protein production, the production load (PL), a summary of all these factors, was utilized.\(^{[32]} \) Therefore, fed-batch cultivations outside of the presented design space were performed at an induction strength of 0, that is, the same CPP combination settings as above for the cultivation temperature and the specific growth rate were used but without induction. These fed-batch cultivations are listed in Table S3, Supporting Information, and the results of the investigation of the PL are presented in Figure S6, Supporting Information.

### 2.2. Intensified Design of Experiments

Two intra-experimental shifts from one CPP set point to another were performed in each fed-batch fermentation to cover three different parameter combinations of the design space within one fermentation. The complete list of all iDoE CPP combination settings and the performed shifts per fed-batch fermentation are provided in Table S2, Supporting Information. These shifts were done in compliance with already published constraints.\(^{[25]} \) Since the inducer was not consumed by the cells, a shift toward lower inducer concentrations was not feasible without heavy and impractical dilution of the fermentation broth. Therefore, the 3D design space was subdivided into three 2D induction planes for the iDoE approach, and shifts were only performed for the temperature and the specific growth rate in the respective induction plane. These shifts were carried out after each theoretically calculated cell doubling, post induction, leading to three phases per fed-batch fermentation. The intensified fed-batch fermentations of the three induction planes were coordinated to guarantee that each CPP combination setting was passed in every phase if overlaid. Figure 1 provides a detailed graphical overview of the operating procedure of the intra-experimental shifts and the performed intensified fed-batch fermentations, shown in the design space and separated in the induction planes.

### 2.3. Data Sets

The static data set derived from an earlier study consisted of 31 fed-batch fermentations (27 CPP combination settings and four replicates) covering the complete design space.\(^{[20]} \) Values for the standard online process parameters were available every minute, while the measurement frequency of the 2D fluorescence probe leads to a value every three minutes. The biomass was measured a single time before induction and then hourly, and the soluble product titer was measured every 2 h from the time point of induction to the last sampling at the end of the process. In total, 589 samples to determine the biomass concentration and 306 samples to analyze the product titer were acquired.

The data set containing the intensified fed-batch fermentations consisted of nine cultivations that were designed to cover the complete design space. The sampling interval and analytical methods for the biomass and soluble product titer analysis were performed as with the static data set, with the sampling interval increased to 30 min after each shift for 2–3 h. In total, 213 samples to determine the biomass concentration and 153 samples to analyze the product titer were acquired. The online available process variables were recorded at the same frequency as that for the static cultivations. The detailed analytical results for the biomass and the soluble product titer of the intensified fed-batch fermentations are provided in Figure S1, Supporting Information. Also, a detailed example of how the CPP changes affect the variables to be modeled and how rapidly these adapt to the new CPP set points is provided in Figure S2, Supporting Information. Further, to exclude a potential memory effect due to the direction of the CPP shifts, the comparison of one experiment, performed reversely to iDoE #3, is presented in Figure S3, Supporting Information.

### 2.4. Data Preprocessing

The data used were stored as Excel spreadsheets with columns representing variables and rows representing observations. Prior to exploratory data analysis and process modeling, every measurement of the available online variables was standardized, along with the time domain, using the z score. This procedure was done to exclude quantitative effects and to specifically account for the change over time. If there was a missing analytical value at a sampling time point for one of the two target variables, the missing value was interpolated using Hermite polynomials, which guaranteed an equally weighted and valid evaluation.

### 2.5. PCA and PARAFAC

PCA and PARAFAC, as described by Bro,\(^{[33]} \) were applied for exploratory data analysis of the online available process data. Both
techniques were performed with MATLAB (2016b, MathWorks, USA) and two freely available toolboxes, N-way [34] and drEEM [35]. PCA was performed on the complete online process data, that is, the standard process variables and the 2D fluorescence data from each approach (DoE and iDoE). PARAFAC was applied to the 2D fluorescence data only. The more detailed and complete comparison of the static and intensified fed-batch cultivations is presented in Figure S5, Supporting Information.

2.6. Hybrid Modeling

2.6.1. Data Sets

Different data sets were used to train the hybrid models:

1) Full-factorial static hybrid model: The first static hybrid model, used as the qualitative reference and derived from an earlier publication, consisted of 25 static fed-batch fermentations (DoE #1, #3–16, #18–20, and #22–27) for model training and 6 static fed-batch fermentations (DoE #2, #4, #9, #17, #21, and #22), including one fermentation each from the duplicate and triplicate runs and three runs chosen by randomization, for model testing (N = 25 + 6).

2) Fractional-factorial static hybrid model: The second static hybrid model was likewise developed as a full-factorial static hybrid model counterpart, but only nine static fed-batch fermentations, the center point, and the corners of the design space, that is, a fractional-factorial design, were used for model training (DoE #1, #3, #7, #9, #14, #19, #21, #25, and #27). This model was developed to allow a comparison between the fractional-factorial static and the iDoE approach with respect to the same amount of input data for model training. The test set contained all static fermentations (DoE #1–27) (N = 9 + 31).

3) iDoE hybrid model: To build the third hybrid model, based on iDoE, all intensified fed-batch fermentations (iDoE #1–9) were considered. To allow a comparison between the full-static and the iDoE hybrid model, the same six static fed-batch fermentations as for the static hybrid model (DoE #2, #4, #9,
4) In addition, for a full comparison of how a model based on iDoE can describe the static design space, a second test set containing all static fermentations was introduced (DoE #1–27) \((\text{N} = 9 + 31)\).

A graphical overview of the respective utilized experiments used for training the three hybrid models is presented in Figure S4, Supporting Information. Hybrid model development and evaluation were accomplished in the stand-alone C# hybrid modeling toolbox (Novasign GmbH, Vienna, Austria), which can be downloaded. Furthermore, the static and the intensified data sets used for modeling are provided as Supporting Information. These are preprocessed and can be used for individual modeling purposes.

2.6.2. Nonparametric Black Box

To predict the values of the response variables, a serial hybrid model structure was chosen. The nonparametric model, an ANN that applies a Levenberg-Marquardt algorithm and is embedded in the hybrid model, was applied to model the known unknowns for the parametric part, that is, the specific growth rate \((\mu, \text{h}^{-1})\) and the soluble product formation rate \((v_{p/x}, \text{gg}^{-1} \text{h}^{-1})\) as propagated predictions.

The ANN had three layers. The nodes of the hidden layer used tangential hyperbolic transfer functions, while the input and output layers used linear transfer functions. There were three inputs: the cultivation temperature (°C), the cumulative inductor mass (mg), and the cumulative feed (L).

2.6.3. Parametric White Box

The hybrid model was developed based on material balances, which were derived for biomass and the soluble product titer assuming an ideally mixed fed-batch reactor. Further, it was assumed that the biomass catalyzes all reactions, therefore, specific rates were used. That is, the estimated rate expressions derived from the nonparametric part are used in the parametric part, as shown in Equation 1 and Equation 2. These equations assume an ideal population, that is, 100 % producing cells and do not consider any emerging subpopulation due to the PL.

\[
\frac{dX}{dt} = \mu \cdot X - D \cdot X \quad (1)
\]

\[
\frac{dP}{dt} = v_{p/x} \cdot X \cdot I_{i/y} - D \cdot P \quad (2)
\]

where \(X\) is the biomass concentration (g L\(^{-1}\)), \(P\) is the soluble product titer (g L\(^{-1}\)), \(I_{i/y}\) is the inductor switch (set to either zero for no induction or one for induction), and \(D\) is the dilution rate (h\(^{-1}\)) to describe the relationship between feed addition (L h\(^{-1}\)) and the reactor volume (L).

2.6.4. Model Validation

The performance of the model with respect to the fit of the experimental data was evaluated using the root mean square error (RMSE) (Equation 3) and the normalized RMSE (NRMSE) (Equation 4). This calculation used the measured value \((\hat{y})\), its estimated counterpart \((\hat{\hat{y}})\) for each sampling point \((t)\), the mean of the measured values \((\bar{y})\), and the total number of observations \((N)\).

\[
\text{RMSE} = \sqrt{\frac{1}{N} \cdot \sum_{i} (y_i - \hat{y}_i)^2} \quad (3)
\]

\[
\text{NRMSE} \% = \frac{\text{RMSE}}{\bar{y}} \cdot 100 \quad (4)
\]

The model was validated using internal cross validation, that is, in the beginning, the hybrid model was derived using the training data. The data were split into training and validation partitions. The training partition was used to build the model, which was then applied to the remaining validation partition. Once no further model improvement was achieved, the model training stopped.

This data partitioning and model development was repeated nine times to account for all possible permutations of eight training and one validation data set. By studying different numbers of nodes, two to eight in steps of one, in the hidden layer of the embedded ANN, the ANN parameters were identified, and four nodes in a single hidden layer were chosen that give the best performance for the fractional-factorial and iDoE hybrid models.

2.6.5. Bootstrap Aggregation

The assessment of the risk of model misprediction based on the random data partitioning during model building used bootstrap aggregation of the individual hybrid models, which can be imagined as model averaging. This averaging of the predictions of multiple models into one gives the operator more control in model selection and represents a robust way to deal with model uncertainties. This approach is similar to a leave-one-batch-out cross validation approach but allows for better control to select individual models of each boot. The bootstrap-aggregated fractional-factorial and iDoE hybrid models each consisted of five individual models, each derived from a different boot, for which the standard deviation (SD) (Equation 5) and the prediction interval (PI) (Equation 6) were calculated to access the model performance:

\[
SD_{i/y} = \sqrt{\frac{1}{n-1} \cdot \sum_{i} (\hat{y}_{\text{boot}(i)} - \hat{y}_{\text{model}(i)})^2} \quad (5)
\]

\[
PI_{i/y} = \hat{y}_{\text{boot}(i)} \pm SD_{i/y} \quad (6)
\]

Therefore, the value of the bootstrap-aggregated prediction \((\hat{y}_{\text{boot}})\), the predicted counterpart from the respective model \((\hat{y}_{\text{model}})\), the index \((i = 1.5)\), and the number of observations for each time point \((n)\) were used. Each generated hybrid model was
Figure 2. Comparative evaluation of the predictive quality of the three developed hybrid models. The model predictions for one exemplary fed-batch fermentation from the test set (DoE #2) are displayed for A) the biomass concentration and B) the soluble product titer. The analytical results (squares), the time point of induction (dashed gray line), and the predictions of the hybrid models (solid lines), including the respective PIs (dashed lines), are indicated: the full-factorial static hybrid model (turquoise), the fractional-factorial static hybrid model (red), and the iDoE hybrid model (green). C) The NRMSE and D) the mean SD of the model predictions for the complete respective test set, using the same color code as above.

derived from a different boot to ensure high generalization ability. This bootstrap-aggregated hybrid model was used for model testing to assess the predictability of the model on new data (external validation) and to investigate the risk of predictive uncertainty.

3. Results

3.1. Setup of the Intensified Design of Experiments

The general operating procedure for the intensified fed-batch fermentations, including the CPP shifts, is presented in Figure 1. Figure 1A indicates the biomass doubling times (generations), the time point of induction, and the performed CPP shifts, that is, the switch to different parameter combinations in the design space, always after one calculated doubling time. Also, the complete design space (Figure 1B) and the more detailed operating scheme of each induction plane are shown, including the location of the starting CPP combination setting for each intensified fed-batch fermentation and the setting after each shift (Figure 1C-E). The intensified experiments were performed so that, if the induction planes are overlaid, each CPP combination setting is characterized in every cell generation.

The comparability of the DoE and iDoE approaches on a cellular level was assessed. The impact of the intra-experimental CPP shifts on the biomass concentration and the soluble product titer is presented in Figure S1, Supporting Information. It was demonstrated that the cells rapidly adapt to new CPP combination settings after a CPP shift (Figure S2, Supporting Information) on the basis of an exemplary iDoE cultivation. To exclude any possible memory effect caused by the CPP shifts, that is, altered behavior with respect to the investigated process variables due to previous CPP combination settings, an additional experiment and exploratory data analysis of the online process data were conducted (Figures S3 and S5, Supporting Information). This exploratory data analysis examined the 2D fluorescence spectra of the static and iDoE approaches as well as the standard online available process variables linked to biomass formation, that is, the cultivation temperature, the accumulated feed, the accumulated inducer, and the base consumption.

3.2. Performance of the Developed Hybrid Models

To investigate and compare the predictive performance of the three developed hybrid models for the biomass and soluble product titer, one exemplary fed-batch fermentation from the test set is presented (Figure 2A,B). This fed-batch fermentation (DoE #2) was chosen since it was present in every test set and was not a replicated cultivation of the training set. Further, the respective
risk of misprediction, that is, the PI, as well as the NRMSE and
the mean SD of the entire test set of each hybrid model, were also
incorporated in this comparative evaluation.

This evaluation revealed that all hybrid models perform on a
comparable level with respect to predicting the biomass. Regarding
the PIs, the full-factorial static hybrid model and the iDoE
hybrid model perform on a comparable level. In contrast, the pre-
dictions from the fractional-factorial static hybrid model are less
precise, displaying broad PIs. A broad PI indicates a high risk
of misprediction, because the different models selected for boot-
strapping are very different in their prediction. This was observed
for both process variables: the biomass concentration (Figure 2A)
and the soluble product titer (Figure 2B). Regarding the NRMSE
(Figure 2C) and mean SD (Figure 2D), a direct comparison shows
that the full-factorial static hybrid model displays the lowest val-
ues for both process variables. The fractional-factorial static hy-
broid model displayed the highest values for both the NRMSE and
the mean SD. The iDoE hybrid model has inferior performance
compared to the full-factorial static hybrid model, but its per-
formance is superior to the fractional-factorial static hybrid model.

3.3. iDoE Hybrid Model Performance on Predicting the Biomass
Concentration

A comprehensive demonstration of the performance of the iDoE
hybrid model using the test set is presented in Figure 3. The iDoE
hybrid model predicted the biomass concentration for all 31 static
cultivations in the scatter plot (Figure 3A) and on nine particular
static fed-batch fermentations, that is, three fed-batch fer-
mentations per induction strength, each performed with one of the
three intended specific growth rates (Figure 3B–D).

The iDoE hybrid model displayed an exceptional ability to
predict the analytical biomass results of the static runs with
high accuracy over the entire cultivation time. With induction
strengths of 0.2 (Figure 3B) and 0.5 (Figure 3C), the model pre-
dictions for all fed-batch fermentations were highly accurate,
including the tightly distributed PIs. At the induction strength
of 0.9 (Figure 3D), higher CPP impacts on the state variables
were observed, impeding the predictions. This was also visible
by the broader PIs. The presented static fed-batch fermentations
DoE #18 and DoE #27 are not perfectly covered by the iDoE hy-
brd model, which displayed a decrease in the biomass concen-
tration. This misprediction was not observed for the third represen-
tative cultivation (DoE #7), for which the biomass trend was
predicted well.

3.4. iDoE Hybrid Model Performance on Predicting the Soluble
Product Titer

As with the biomass concentration, the performance of the iDoE
hybrid model for predicting the soluble product titer of all 31
static cultivations is showcased in Figure 4. To obtain a consistent
Figure 4. Performance of the bootstrap-aggregated iDoE hybrid model in predicting the soluble product titer of the 31 test cultivations. A) The scatter plot of the hybrid model on the training data (orange dots) and the test data (cyan triangles). The model predictions for the individual fed-batch fermentations are displayed for each induction strength. For B) I 0.2, C) I 0.5, and D) I 0.9, the analytical results (symbols), the respective prediction (solid lines), and the PI (dashed lines) of the iDoE hybrid model are indicated. The IDs of the presented fed-batch fermentations are listed.

impression of the iDoE model quality, the cultivations shown are the same as those in Figure 3.

The iDoE hybrid model was able to predict the soluble product titer of the full-factorial design space with adequate accuracy, as presented for the individual induction strengths. For the induction strength of 0.2 (Figure 4B), accurate predictions were observed for two out of three cultivations. Only cultivation DoE #3 displayed an overestimation of the analytical values. Similar behavior was also observed for one cultivation that was performed with an induction strength of 0.5 (Figure 4C), namely DoE #13, while predictions of the other two cultivations matched the analytical values and displayed small PIs. Similar results were obtained for a strength of 0.9 (Figure 4D). The trend for two cultivations was not predicted completely well (DoE #7 and DoE #18), while the remaining cultivation was predicted accurately.

A complete comparison of the predictive quality of the static and intensified bootstrap-aggregated hybrid models, applied on the respective test sets, is presented in Table 1 using the $R^2$, RMSE, and NRMSE as criteria for model comparison.

| Table 1. Model comparison of the developed bootstrap-aggregated static hybrid models and the bootstrap-aggregated iDoE hybrid model. The results of the three models applied to the respective test sets are presented, including the respective $R^2$, RMSE, and NRMSE, all rounded to two decimal places, and the required number of experiments. The number of fed-batch fermentations used for model training and model testing is indicated in brackets. |
|---|---|---|---|---|
| Target variable | Full-factorial static hybrid model ($N = 25 + 6$) | iDoE hybrid model ($N = 9 + 6$) | Fractional-factorial static hybrid model ($N = 9 + 31$) | iDoE hybrid model ($N = 9 + 31$) |
| $R^2$ | 0.98 | 0.98 | 0.97 | 0.97 |
| RMSE [g L$^{-1}$] | 1.10 | 1.12 | 1.29 | 1.19 |
| NRMSE [%] | 5.77 | 5.86 | 6.83 | 6.30 |
| No. of experiments | 31 | 9 | 9 | 9 |

© 2020 The Authors. Biotechnology Journal published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
4. Discussion

The evaluation of the iDoE approach (Figure 1) on a full-factorial design space and the quality of the generated data sets was of primary interest in this study. The investigation into the comparability of the DoE and iDoE setups on a cellular level was crucial and of high interest for the consecutive modeling steps. Applying the described intra-experimental shifts, it was demonstrated that the cells can cope with the iDoE setup and the easy to express recombinant human superoxide dismutase. We showed that cells are able to rapidly adapt to new process conditions within 1 h after the change of conditions (Figure S2, Supporting Information) and, regarding the outcome, that the shift direction does not matter (Figure S3, Supporting Information). Hence, the defined intervals between the changes of process conditions were not too short, providing cells adequate residence times for each CPP setting. Moreover, the exploratory data analysis of the DoE and iDoE data in Figure S5, Supporting Information, displayed no significant differences on the cellular and process level between both approaches. These results also strongly support the assumption that appropriate CPP shifts do not provoke persistent cellular memory effects. In addition to information on dynamics in response to changes, iDoE data displays similar information content as data from conventional static experiments, which can be seen as sound basis for more detailed data interpretation via modeling approaches. However, the general usability of the iDoE approach must be investigated, for example, for more input factors, a variety of target proteins with different characteristics, for example, cytotoxicity, and especially the applicability on other organisms.

From a modeling and prediction perspective, hybrid models utilizing either iDoE or DoE data were evaluated with respect to their prediction performance. The hybrid model established with iDoE data was able to predict the biomass for the test data set (Figure 3), containing all static DoE cultivations ($N = 31$), with an accuracy similar to the static full-factorial hybrid model. There was also good accordance with the analytical error of the biomass determination (3.6%). The prediction performance for soluble product titer was on an acceptable level and again comparable to predictions with the full-factorial hybrid model (Figure 4). In general, it also has to be kept in mind that the test set for the full-factorial static hybrid model consisted of only six cultivations which were used to calculate the RMSE, NRMSE, and mean SD. Therefore, its performance in comparison to the iDoE hybrid model, using 31 cultivations in the test set, should not be overrated. We also verified the potential of the idea to save time and costs simply by reducing the number of static experiments. Therefore, a fractional-factorial data set, comprising only the center point and the corners of the static DoE, was used for model building (Figure S4, Supporting Information). This approach resulted in a model with significantly reduced prediction quality and a strongly increased model uncertainty (Figure 2). The results, summarized in Table 1, clearly demonstrate the superiority of iDoE data which is most probably based on a significantly increased information entropy, as every single iDoE experiment contains data from three different CPP settings. With the prospect for process control, accurately modeling the response to a variable that changes over time is a great advantage. Even though the inputs to the ANN, that is, the three chosen CPPs, do not fully describe all possible process responses, these are easily controllable, thus enabling model predictive control applications in the future.

Further, the somehow limited prediction performance for the soluble product titer observed for all hybrid models built in this study is most likely caused by the formation of nonproducing subpopulations during the induction phase of the process. In E. coli cell banks, the presence of a small population of plasmid-free cells is a known phenomenon and even application of selection pressure along the production process rendered to be of limited efficiency in suppressing this subpopulation during the production phase. As we assume a homogeneous population of producer cells, in Equation 2, this decoupled the biomass from the product formation. The key problem in this context is that there is no information on the distribution of producer and nonproducer in our datasets. There is no analytical method available that facilitates differentiation between producing and nonproducing subpopulations with the required accuracy or, for example, without introducing an additional fluorescent protein, which is not applicable for industrial production processes. As the load level, triggered by product formation, directly impacts the difference in growth rates of these subpopulations, we introduced the PL concept. The obtained PL values (up to 30%), presented in Figure S6, Supporting Information, were in good accordance with the reference literature and, further, are reasonable from a metabolic point of view, that is, an increase in the induction strength, as well as the cultivation temperature, raised the PL. We are aware that the pure description of the PL is not the solution to the occurring limitations of accurately predicting the soluble product titer but rather a starting point to this multidimensional restriction of the model performance, which we are not able to fully explain. However, this limitation will remain at the moment, since a precise analytical method to approach this nonproducing population is not available. As the limiting problem for more accurate predictions is known, we anticipate a predictive improvement by incorporating a suitable term in the white box, that is, solely taking the producer population for the product formation into account and not the entire predicted biomass. This again highlights the advantages of knowledge incorporation using hybrid models. However, for systems with constitutive product formation, and therefore without selection pressure introduced by induction, the predictability of the product is assumed to be much higher.

In conclusion, the concept of performing iDoE is rather new in upstream processing, and it has never before been tested in a comprehensive comparison between static and dynamic cultivations of a full-factorial design space. Besides the more common considerations, for example, which process variables will be analyzed and which data will be measured, a major and more conceptual matter must be considered when an iDoE is designed: the conduction of the intra-experimental CPP shifts. Adequate time for the adaption of the cells in each phase, the number of shifts per cultivation, and a reasonable direction of all shifts and their magnitudes are highly important in generating meaningful process data.

Further, we do not propose iDoE for the discovery of an unknown design space where particular CPP combinations can lead to irreversible cell damage. If cells do not recover after a shift, the following learning rate of the hybrid model might be very low and inaccurate. Additionally, the number of required cultivations was
significantly reduced compared to the full-factorial static hybrid model, that is, the application of iDoE in this study led to an acceleration of the process characterization time by more than 66%, while the analytical effort increased by 20% due to the temporal higher sampling frequency. This is also highly valuable from the economical aspect to keep up in terms of budget restrictions, that is, saving working hours and raw materials for the preparation and execution of the fermentation of two thirds of the complete experimental setup.

Furthermore, the iDoE provided a lot of information about the behavior of cells to changes in the process which enabled the development of a hybrid model with a good generalization ability. Since a lot of information was gathered on how cells react on process changes, utilizing the iDoE hybrid model structure for advanced process control is the logical next step. The great potential for using iDoE in combination with hybrid models to speed up process characterization, as was demonstrated in this work, is of high interest for the biopharmaceutical industry, for example, to keep up with timelines, budget restrictions, and return on investments.

Supporting Information
Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements
The authors would like to thank the Austrian Research Promotion Agency (FFG) for their support (Research Studio Austria, 859219). The authors would like to thank Roger Dalmou Diaz (University of Natural Resources and Life Sciences, Vienna) for developing the prototype of the Novasign Hybrid Modeling Toolbox and Lina Vranitzky for her support during the fed-batch cultivations and for conducting the ELISA measurements. The authors would also like to thank Moritz von Stosch and Michael Melcher for critical review and input during the preparation of this manuscript.

Conflict of Interest
Gerald Striedner and Mark Dürkop hold shares of Novasign GmbH.

Keywords
machine learning, process control, quality by design

Received: March 16, 2020  
Revised: May 11, 2020  
Published online: June 22, 2020

[1] A. S. Rathore, H. Winkle, Nat. Biotechnol. 2009, 27, 26.
[2] V. Kumar, A. Bhalia, A. S. Rathore, Biotechnol. Prog. 2014, 30, 86.
[3] L. Zhang, S. Mao, Asian J Pharm. Sci. 2017, 12, 1.
[4] T. Lundstedt, E. Seifert, L. Abramo, B. Thelin, A. Nyström, J. Pettersen, R. Bergman, Chemom. Intell. Lab. Syst. 1998, 42, 3.
[5] K.-M. Lee, D. F. Gilmore, Appl. Biochem. Biotechnol. 2006, 135, 101.
[6] S. J. Kalil, F. Maugeri, M. I. Rodrigues, Process Biochem. 2000, 35, 539.
[7] M. S. Tanyildizi, D. Özer, M. Elibol, Process Biochem. 2005, 40, 2291.
[8] F. Torkashvand B. Vaziri, S. Maleknia, A. Heyrdari, M. Vossoughi, F. Davarni, F. Mahboudi, PloS One 2015, 10, e0140597.
[9] G. Q. Liu, X.-L. Wang, Appl. Microbiol. Biotechnol. 2007, 74, 78.
[10] C. F. Mandenius, R. Gustavsson, J. Chem. Technol. Biotechnol. 2015, 90, 215.
[11] J. Claßen, F. Aupert, K. F. Reardon, D. Solle, T. Scheper, Anal. Bioanal. Chem. 2017, 409, 651.
[12] S. Craven, J. Whelan, B. Glennon, J. Process Control 2014, 24, 344.
[13] R. A. Harshman, M. E. Lundy, Comput. Stat. Data Anal. 1994, 18, 39.
[14] J. Shlens, Int. J. Remote Sens. 2014, 51, 1.
[15] P. Kadlec, B. Gabrys, S. Strandt, Comput. Chem. Eng. 2009, 33, 795.
[16] L. Mears, S. M. Stocks, M. O. Albaek, G. Sin, K. V. Gernaey, Trends Biotechnol. 2017, 35, 914.
[17] M. von Stosch, S. Davy, K. Francois, V. Galvanauskas, J.-M. Hamelink, A. Luebbert, M. Mayer, R. Oliveira, R. O’Kennedy, P. Rice, J. Glassey, Biotechnol. J. 2014, 9, 719.
[18] B. D. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, Oxford, UK 1996.
[19] B. Bayer, B. Sissolak, M. Duerkop, M. von Stosch, G. Striedner, Bioprocess Biosyst. Eng. 2020, 43, 169.
[20] B. Bayer, M. Von Stosch, G. Striedner, M. Duerkop, Biotechnol. J. 2020, 15, 1900551.
[21] S. Varma, R. Simon, BMC Bioinformatics 2006, 7, 91.
[22] J. Pinto, C. R. de Azevedo, R. Oliveira, M. von Stosch, Bioprocess Biosyst. Eng. 2019, 42, 1853.
[23] M. von Stosch, J. M. Hamelink, R. Oliveira, Biotechnol. Prog. 2016, 32, 1343.
[24] T. Patarinska, D. Dochain, S. N. Agathos, L. Ganovski, Bioprocess Eng. 2000, 22, 517.
[25] M. von Stosch, M. J. Willis, Eng. Life Sci. 2017, 17, 1173.
[26] B. Bayer, M. von Stosch, M. Melcher, M. Duerkop, G. Striedner, Eng. Life Sci. 2020, 20, 26.
[27] M. Cserjan-Puschmann, W. Kramer, E. Duerrschmid, G. Striedner, K. Bayer, Appl. Microbiol. Biotechnol. 1999, 53, 43.
[28] T. Porstmann, R. Wietschke, H. Schmechta, R. Grunow, B. Porstmann, R. Bleiber, M. Pergande, S. Stachat, R. von Baehr, Clin. Chim. Acta 1988, 171, 1.
[29] K. Marisch, K. Bayer, M. Cserjan-Puschmann, M. Luchner, G. Striedner, Microbiol. Cell Fact. 2013, 12, 58.
[30] M. Luchner, G. Striedner, M. Cserjan-Puschmann, F. Strobl, K. Bayer, J. Chem. Technol. Biotechnol. 2015, 90, 283.
[31] M. Melcher, T. Scharl, B. Spangl, M. Luchner, M. Cserjan, K. Bayer, F. Leisch, G. Striedner, Biotechnol. J. 2015, 10, 1770.
[32] P. Rugbjerg, N. Myling-Petersen, A. Porse, K. Sarup-Lytzen, M. O. A. Sommer, Nat. Commun. 2018, 9, 787.
[33] R. Bro, Chemom. Intell. Lab. Syst. 1997, 38, 149.
[34] C. A. Andersson, R. Bro, Chemom. Intell. Lab. Syst. 2000, 52, 1.
[35] K. R. Murphy, C. A. Stedmon, D. Graeber, R. Bro, Anal. Methods 2013, 5, 6557.
[36] D. A. Freedman, Ann. Stat. 1981, 9, 1218.
[37] A. Schuller, M. Cserjan-Puschmann, C. Tauer, J. Jarmer, M. Wagnerknecht, D. Reinisch, R. Grabherr, G. Striedner, Microb. Cell Fact. 2020, 19, 58.