von Stosch M, Hamelink J-M, Oliveira R.

[Toward intensifying design of experiments in upstream bioprocess development: An industrial *Escherichia coli* feasibility study.](#)

Biotechnology Progress (2016).

DOI: 10.1002/btpr.2295

Copyright:

This is the peer reviewed version of the above article, which has been published in final form at <http://dx.doi.org/10.1002/btpr.2295>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

Date deposited:

04/05/2016

Embargo release date:

29 April 2017

1 Towards intensifying Design of Experiments in upstream
2 bioprocess development: An industrial *E. coli* feasibility study

3

4 Moritz von Stosch* (1,2), Jan-Martijn Hamelink (3,4), Rui Oliveira (2)

5 1 – CEAM, Faculty of Science, Agriculture and Engineering, Newcastle University, Newcastle

6 upon Tyne, NE1 7RU, UK

7 2 - REQUIMTE/DQ, Faculty of Science and Technology, University Nova de Lisboa, Campus de

8 Caparica, 2829-516 Caparica, Portugal

9 3 - GSK Vaccines, Laval, Canada

10 4 – Current affiliation: Lallemand Inc., Montreal, Canada

11

12 *Corresponding author

13   **Abstract**

14   In this study, step variations in temperature, pH and carbon substrate feeding rate were
15   performed within five high cell density *E. coli* fermentations to assess whether intra-experiment
16   step changes, can principally be used to exploit the process operation space in a design of
17   experiment manner.

18   A dynamic process modeling approach was adopted to determine parameter interactions. A
19   bioreactor model was integrated with an artificial neural network that describes biomass and
20   product formation rates as function of varied fed-batch fermentation conditions for
21   heterologous protein production.

22   A model reliability measure was introduced to assess in which process region the model can be
23   expected to predict process states accurately. It was found that the model could accurately
24   predict process states of multiple fermentations performed at fixed conditions within the
25   determined validity domain. The results suggest that intra-experimental variations of process
26   conditions could be used to reduce the number of experiments by a factor, which in limit would
27   be equivalent to the number of intra-experimental variations per experiment.

28   **Keywords**

29   PAT, QbD, upstream bioprocess development/optimization, Intensified Design of Experiments,
30   Hybrid modeling, dynamic excitations,

31 **Introduction**

32 In the Process Analytical Technology (PAT) initiative and Quality by Design (QbD) paradigm the
33 focus is on achieving quality products through online process monitoring and control instead of
34 off-line testing [1, 2]. Understanding how process parameters and raw material attributes affect
35 product quality attributes is an essential precondition [3-7]. Risk-based approaches that use
36 prior available knowledge about process and product are typically employed in a first step to
37 determine the most critical process parameters (CPPs) and critical material attributes (CMAs) as
38 well as the most critical product quality attributes (CQAs) [1, 3]. In the subsequent step,
39 correlations between critical parameters and attributes are evaluated using design of
40 experiment (DoE) methods, where levels of CPPs and CMAs are systematically varied in a set of
41 experiments [3, 6-8]. In each individual experiment, a fixed combination of CPPs and CMAs is
42 investigated and the experimentally obtained responses in CQAs are subsequently analyzed,
43 typically using Multivariate Data Analysis (MVDA) [3, 5-9].

44 Application of this DoE-MVDA approach can be challenging and laborious in particular for
45 upstream bioprocess development (e.g. microbial fermentation) because of the complexity of
46 the process and its sensitivity to many parameters [5, 10]. For instance, for an increasing
47 number of CPPs and CMAs, the number of experiments might become impractical [10].

48 Moreover, interactions between CPPs and CMAs are complex and nonlinear, and require more
49 complex experimental designs . Significant reductions in cost and more importantly in process
50 development timelines could be obtained, if it was possible to gain the same degrees of process
51 understanding using a reduced set of experiments.

This paper studies the idea of applying intra-experimental variations of CPPs instead of fixed values during each experiment, herein referred to as intensified Design of Experiments (iDoE). This is a new concept in upstream bioprocess development and optimization, but excitation of process conditions in a dynamic manner (step-changes, ramps, oscillations) is a well-known and studied subject in process control for model and system identification [11]. For instance, Akesson et al. [12] or Schaepe et al. [13] use dynamic changes in the substrate feeding of *E.coli* fermentations to determine the maximal possible feeding rate, whilst avoiding an overflow metabolism. In a similar way, Dietzsch et al [14] applied pulses in the methanol feeding of *Pichia pastoris* cultivations, to determine methanol uptake for model based feeding optimization. However, to the best of our knowledge systematic intra-experiment variation of multiple control variables (CPPs) in upstream bioprocess development has not been reported. Intensifying the DoE process has the potential to significantly reduce the number of experiments, and therefore costs and timelines.

Execution of dynamic variations in CPPs results in intra-experiment variations in CQA responses, which requires an alternative analysis method to the traditional MVDA approach to determine parameter interactions. Either CQA responses have to be converted into time-independent response variables for the analysis [15] or a dynamic process modeling approach need to be applied. It has been shown that dynamic models can efficiently be developed by integrating knowledge from first-principles with data-derived approaches into so-called hybrid models [16, 17]. A general framework for their development has been reported [18]. Hybrid models can potentially reduce experimental load, increase prediction performance and improve process understanding [19, 20].

In this study, an industrial microbial fermentation process with heterologous protein production is investigated. The objective was to investigate whether a model that was developed using experiments with dynamic variations in CPPs could be used to describe experiments at constant CPPs, as this would allow characterizing the design space in a more efficient way. Within a reduced number of experiments, multiple systematic changes of input process variables are applied with the goal to explore the experimental design space in an accelerated manner. The output responses, biomass formation and productivity, are then correlated with the varying process parameters, employing a dynamic hybrid model. A clustering approach is applied to determine the validity domain of the developed model. Finally, model performance is tested with process data of multiple batches that were performed at constant process conditions within the validity domain.

Methods

Intensified Design of Experiments

A target protein of interest is expressed in *E.coli* in a high-cell density fermentation process that consists of three phases: 1) growth in batch mode, 2) growth in fed-batch mode; and 3) induction of product formation by addition of IPTG (Isopropyl β -D-1-thiogalactopyranoside). Five iDoE batches were performed to investigate the effect of induction parameters on biomass and product formation. All five batches started with identical batch volume, were induced at a biomass concentration around OD 40 and were harvested 30 hours after. Conditions were varied over the course of the induction phase by automated bioprocess control. Process temperature, pH and feed rate were scaled around the center point (q) of the design in Figure

1B ($T=0$; $pH=0$ and $u_c=0$) to preserve process confidentiality. Figure 1A shows how set-points in four iDoE batches (I to IV) were changed to cover a 2 factor Doehlert design space for induction temperature (5 levels ranging from -1.5 to 0.5) and induction feed rate (3 levels ranging from 0.1 to 1.9), while induction pH was kept constant at 1.0. Set-points were counter-rotated within the 2 dimensional design space to assure coverage of conditions during early stage as well as late stage of the induction phase (i.e. time-sensitive behavior). In addition, these counter-rotated experiments allow to assess the sensitivity of the process with respect to changes in the magnitude and direction of set-point changes. The induction conditions of the fifth iDoE batch (Figure 1B: batch V) range across a 3 dimensional design space for temperature (ranging from -1.0 to 1.0), pH (ranging from -1.0 to 1.0) and carbon substrate feed rate (ranging from -1.0 to 1.0).

Additional to the iDoE investigation, the same set of experimental conditions shown in Figure 1A were examined under traditional fixed parameter fermentations. These batches were designated the letters [a] to [g] of which the center point was conducted in duplicate: [g₁] and [g₂]. These fixed condition batches were induced around OD 40.

An additional 23 fixed-condition batches were performed within the experimental space represented by Figure 1B (points not shown), covering a 4 dimensional design space for induction temperature, feed rate, pH and induction OD (ranging from OD 40 to 80, which is not shown in Figure 1B). Within this design space, one batch was induced at OD 40 employing the center point conditions: batch [q].

115 All fixed condition experiments were part of a process development study to find optimal
116 induction conditions for recombinant protein expression, and that comprised multiple design
117 spaces. Results of this study have been reported in [20]. Process conditions of the five iDoE and
118 39 fixed condition batches are presented in Table 1. Some constant condition experiments
119 overlap with the design space in Figure 1A and are included in this figure: batches [h] to [p].

120 Responses to set-point changes were different for each of the induction conditions due to
121 differences in control loops. Delay times for gravimetric feed control, temperature control
122 (water jacket control using heat exchange with steam and chilled water), and one-way pH
123 control (with a base solution) were in the order of 5 sec, 15 min and 5 min, respectively. Set-
124 point changes towards a more acidic pH were achieved using unregulated pH decline
125 (metabolically driven), at a rate of approximately -0.3 scaled pH units/hour.

126 A small number of in-process samples were taken at pre-induction and over the course of the
127 induction phase, up to a maximum of six samples per batch. Biomass concentration was
128 determined by measurement of optical density (OD) at 650 nm. Quantification of the soluble
129 cytoplasmic expressed product was performed by densitometry of Coomassie stained SDS-
130 PAGE and by RP-HPLC analysis on separated fractions of mechanically lysed cells.

131 High frequency on-line process data (temperature, pH and feed and base balance readings) was
132 averaged in 30 minute intervals to filter noise and high-frequency variations. Since the process
133 was under control, high-frequency variations in the online-measurements are mostly due to
134 noise, wherefore little information about the process is lost through averaging. The 30 minute
135 interval was chosen by visual inspection of the signals. Averaging increases the hybrid model

identification efficiency whilst it does not reduce applicability since the modeling focus is on biomass growth and product formation, which exhibit rather slow dynamics as opposed to the fast dynamics of e.g. the dissolved oxygen concentration.

Experimental results were used to develop a dynamic hybrid model in which a bioreactor model is integrated with an artificial neural network that describes biomass and product formation rates as functions of the varying induction conditions.

Dynamic Hybrid Process Model

For the described fed-batch *E. coli* process, the material balances provide a sound and generally valid modeling framework. The balance equations for biomass and product were derived assuming an ideally mixed reactor, with X and P designating biomass and product concentrations, respectively. The balances for biomass and product formation are written using specific rates:

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

$$\frac{dP}{dt} = v_P \cdot X - D \cdot P \quad (2)$$

where D is the dilution rate equal to $D = (u_{Feed} + u_{Base})/V$ in fed-batch mode, μ is the specific biomass growth rate and v_P is the specific product formation rate. The product is expressed in the cytoplasm and remains within the cells. When analyzing the productivity of *E. coli* cultures the specific productivity is typically regarded as an important measure [21, 22]. It is useful to introduce an equation for specific productivity, P/X :

$$\frac{d(\frac{P}{X})}{dt} = v_{P/X} \cdot I \quad (3)$$

156 where the specific rate $v_{P/X}$ describes the change of the specific productivity along time and I
 157 is the induction parameter (0 before induction, 1 after induction). By use of the chain rule,
 158 combining equations (1) and (3) and rearranging, it can be shown that the specific product
 159 formation rate $v_P = v_{P/X} \cdot I + \frac{P}{X} \cdot \mu$.

160 A correlation between biomass growth and base addition was used to exploit the available
 161 online base consumption data and to compensate for the relative low number of biomass data
 162 points to capture the dynamics of biomass growth, i.e.:

$$163 \quad \int_{t_0}^t u_{Base} \cdot dt = a_{Base/X} \cdot (X(t) \cdot V(t) - X(t_0) \cdot V(t_0)) \quad (4)$$

164 where u_{Base} is the gravimetric base addition rate and $a_{Base/X}$ is a function (which depends on
 165 pH and temperature) correlating base consumption and biomass growth.

166 The specific rates, μ and $v_{P/X}$, as well as the correlation coefficient $a_{Base/X}$, are considered
 167 functions of the process conditions, such accounting for step-changes in pH, temperature and
 168 carbon feeding rate. They are not directly measurable or derivable and in this case they were
 169 modeled by means of an artificial neural network as described in the following section.

170 **Modeling of rates by neural networks**

171 The specific rates and base correlation coefficient were approximated by two artificial neural
 172 networks (ANNs), each comprising three layers. The nodes in the input and output layers of
 173 both networks were chosen to have linear transfer functions, whereas the nodes in the hidden
 174 layer have tangential hyperbolic transfer functions, $g(\cdot)$ and $h(\cdot)$, i.e.:

$$175 \quad [\mu, v_{P/X}] = w_2 \cdot g(w_1 \cdot x_1 + b_1) + b_2 \quad (4)$$

$$[a_{Base/X}] = w_4 \cdot h(w_3 \cdot x_2 + b_3) + b_4 \quad (5)$$

Where w_1 to w_4 are the weights of the connections between the nodes of the networks, b_1 to b_4 the biases and x_1 and x_2 the vectors of inputs. The performance of different network inputs was tested and the variables with the highest functional dependence were included as neural network inputs. Vector x_1 consists of five variables: X , P/X , T , pH and the carbon substrate feeding rate, u_C , which is the product of u_{Feed} and the carbon substrate concentration in the feed solution. The vector x_2 comprises two variables: T and pH . Due to the differences in the input domains, two neural networks were utilized decoupling the estimation of $a_{Base/X}$ from biomass concentration, specific productivity and the carbon substrate feeding.

The networks were trained on data of three iDoE experiments: II, IV and V (Figure 1), referred to as the training data set, minimizing the weighted residual of the model estimates, c , and the experimental data points c_{exp} :

$$\min_w \left\{ \sum \frac{(c_{exp} - c(t,w))^2}{\sigma_c} \right\}, \quad (6)$$

where σ_c is the variance calculated from all experimental values of the training data for each concentration. The gradient based MATLAB "*lsqnonlin*" function was used for the minimization and analytic gradients were provided employing the sensitivity method [17, 18]. Due to the possible convergence to local minima, the parameter identification was repeated 40 times from random initial weight values to find parameters that approximate the data well. The sensitivity equations were integrated along with material balances (1) and (2) using an Euler forward integration method with a fixed step size of 0.3 hours, which was found to be appropriate with

respect to the model and parameter identification performance in an initial test. Experimental values were used as initial values for the material balances, zeros were used in case of the sensitivities equations (as the initial values are independent of the parameter values). The data of a fourth iDoE experiment (Figure 1: I), referred to as the validation data set, was used to determine the point at which the training is stopped, i.e. simple cross-validation [18]. The fifth iDoE experiment (Figure 1: III) was used to test the performance of the model, referred to as the test data set. This batch was purposely chosen as test batch, since it had shown to have a significant negative effect on model performance if selected in training or validation sets. Treated as independent data set, its particular outlier behavior does not deteriorate model performance and can be investigated more in detail.

Different network topologies were studied and their performance was compared using the negative Bayesian Information Criteria (BIC) [23]. The best structure was chosen as the one with the greatest BIC value for the validation set (the one closest to zero), whilst showing consistent performance also for the training set: i.e. three nodes in the hidden layer of the rate estimating ANN and two nodes in the neural network describing the coefficient correlating base consumption and biomass growth.

Modeling performance of different experiments and variables was assessed using a weighted mean least squared error criterion:

$$MSE = \frac{1}{n_D} \cdot \sum_{n_D} \frac{(c_{exp} - c(t, w))^2}{\sigma_c}, \quad (7)$$

215 with n_D the number of data points of each variable. This criterion provides an intuitive
216 measure: performance increases with decreasing MSE value.

217 **Model Validity Domain**

218 Hybrid models have been reported to exhibit good extrapolation properties, due to
219 incorporated fundamental knowledge [16, 18, 19, 24]. The limiting component for
220 extrapolation is the incorporated data-driven model. Neural network prediction performance
221 will deteriorate when confronted with input values different to training data [16, 25]. The
222 training data frames the experience domain of the neural network, i.e. the domain in which the
223 network is “experienced” and therefore expected to perform well. A model validity measure
224 can be used to assess the expected performance of test data. Two measures have been
225 reported in hybrid modeling literature, namely a convex-hull approach and a clustering
226 approach [16, 25, 26]. In the first approach a convex hull is constructed on the input training
227 data, defining the validity domain irrespective of the internal data distribution. In case of
228 unevenly distributed input data, this measure might be overly optimistic [16, 26]. The clustering
229 approach is an alternative that was used in this study, since it can account for unevenly
230 distributed data. Input data from different experimental regions is regrouped into so-called
231 clusters after which model validity for a new data point can be determined by its distance to
232 the nearest cluster [16, 25].

233 For a new data point $x_1(\tau)$ the distance to the nearest cluster is compared to the cluster width,
234 which is used to quantify the degree of membership of input vector $x_1(\tau)$ to the network
235 domain of experience at some time τ , i.e.: $m(x_1(\tau), c_{n_C}, \sigma_{n_C})$, the membership function. The

function was constructed by clustering the input training data of the ANN, $x_{1,Train}(\tau)$, using n_C radial basis function clusters, which positions, c_{n_C} , and width, σ_{n_C} , were identified by the k-means algorithm (see Teixeira et al [25] for more details).

The greater the distance between inputs $x_1(\tau)$ and the training inputs $x_{1,Train}(\tau)$ the lower the membership value (maximum 1, minimum 0). With decreasing membership values the risk of mispredictions, $r(\tau) = 1 - m(x_1(\tau), c_{n_C}, \sigma_{n_C})$, grows. The time averaged risk, R , provides a measure of the risk of misprediction for a complete fermentation of duration t_B , i.e.:

$$R = \frac{1}{t_B} \cdot \int_0^{t_B} r(\tau) \cdot d\tau = \frac{1}{t_B} \cdot \int_0^{t_B} (1 - m(x_1(\tau), c_{n_C}, \sigma_{n_C})) \cdot d\tau \quad . \quad (8)$$

Validation and test data can be used to understand how critical certain time averaged risk levels are with regard to model reliability.

Results and Discussion

Analysis of hybrid model estimation performance

The hybrid model regression plots in Figure 2 illustrate substantial agreement between measured and model estimated biomass, product concentrations and specific productivity for the training (experiments II, IV and V) and validation (experiment I) sets. Absence of overnight sampling resulted in a gap in biomass data between OD 95 and 130; however the high-resolution base consumption data supports the model in capturing the biomass growth kinetics in this region. The agreement between the correlation for biomass formation and accumulated base addition affirms this.

Training and validation batches cover a large part of the domains presented by Figure 1, and in general, we can expect the model to accurately predict biomass and product formation. The hybrid model was tested against batch data of experiment III. While biomass measurements and estimates are in good agreement, productivity regression is deteriorating with increasing volumetric and specific productivity (green x in Figure 2). Besides the fact that iDoE III is going through a region not fully covered by training and validation data, it also caused similar reduced performance when tried with other combinations of training and validation data or model structures. How model validity domain and discrepancies between model performance and experimental results are related will be further analyzed in the following sections.

Analysis of the validity domain

Validity domain definition

The ANN embedded in the hybrid model can be expected to be the limiting component for the accuracy of model predictions, in cases where input values are different to training data. In this respect, a membership function was defined to quantify the degree to which a new set of input values is similar to training data (see section Model Validity Domain). Contour plots of the membership function are presented in Figure 3, as function of the five ANN input variables, X , P/X , T , pH and u_C . The values of this experience measure range from 0 – not a member to 1 – a member. The systematic variations in process conditions, i.e. in pH, temperature and feeding rate, form the shape of the membership functions. It must be noted that although contour plots are helpful visual tools, they only show a 2-dimensional result of the validity domain that is not completely representative for the 5-dimensional domain that we are studying.

276 Time dependent risk profiles were constructed from the membership function: $r(\tau) = 1 -$
277 $m(x_1(\tau), c_{n_C}, \sigma_{n_C})$. Figure 4 shows dynamic risk profiles of iDoE batches I to V and selected
278 fixed condition batches. Time-averaged risk values of all performed batches are included in
279 Table 1.

280 Low time averaged risk values in the order of 10^{-3} were obtained for training data (iDoE batches
281 II, IV and V). For the input values of the validation data (iDoE batch I), risk values vary between
282 0.1 and 0.65 over the course of the fermentation, which corresponds to a time averaged risk
283 value of 0.35. The model performance for the four training and validation batches was
284 demonstrated by significant agreement between measured and estimated biomass and product
285 formation. Based on the range of time averaged risk values for training and validation data ($R=0$
286 to 0.4) and the respectable model performance for these sets, we hypothesize that model
287 performance can be generally assumed to be good for averaged risk values < 0.4 . This will also
288 be verified in the following paragraphs.

289 **Validity domain verification**

290 The developed hybrid model was applied to fixed condition experiments outlined in Table 1.
291 The batches were divided in two groups: a group that yielded averaged risk values below 0.4
292 and a group of experiments with greater risk values (Table 1). For the group of fixed condition
293 batches with $R < 0.4$ the temperature ranges between -1.5 and 0.3 and substrate feeding rate
294 between 0.1 and 2.0, whereas the pH is almost identical across the experiments (ranging from
295 pH 0.9 to 1.0). The iDoE batches were all induced around OD 40, but some higher induced fixed
296 condition batches (OD 50 to OD 70) generated low risk values ($R < 0.4$) as well, suggesting model

reliability in this domain of induction at higher biomass. The low risk batches were each designated a unique batch number: [b] to [p] in Table 1 (Note that the letter l was not used as label due to its similarity to the iDoE label l). Experiments [b] to [g] correspond to the batches in domain Figure 1A. The wide-ranging coverage of the iDoE batches in this domain account for their low risk values. Experiment [a] however, although within this 2 factor Doehlert design space, yields a time averaged risk above 0.4. The reason is that the training data of experiment II includes the conditions of experiment [a] only for the first 4 hours of the 30 hour induction period. As a consequence, the risk of model misprediction for experiment [a] is low during the first four hours of induction, but much higher beyond this point (see Figure 4), leading to $R > 0.4$.

In Figure 5 it can be observed that model performance for biomass growth and base consumption decreases for averaged risk values above 0.4. For productivity predictions, the same trend holds, but three outliers with risk values below 0.4 can be observed, namely experiments III, [c] and [k]. Experiments [c] and [k] were carried out in a domain of low temperature, high feed rate and at pH 1.0, while iDoE experiment III was conducted at similar conditions during the first 8 hours of induction. These three batches generated productivity levels significantly higher than the model predicted. However, several other experiments, conducted with conditions in the same domain region were accurately described by the model: i.e. batches [d], [m], [o1] to [o4] and [p]. This suggests that the behavior of product formation within this region is not completely captured by the model.

Figure 5 also shows several batches with $R > 0.4$, but with respectable model performance for both biomass growth and product formation. Particularly experiments performed within the

space presented by Figure 1B, but induced over a large range of biomass concentrations (OD 40 to 80) were described well by the model: e.g. experiment [q], center-point, induced at approximately OD 40. The model validity region could have been possibly augmented by performing an iDoE experiment across this large 4 dimensional space (T, pH, u_c and induction OD).

Analysis of the prediction performance

Time profiles of biomass and specific product concentration are shown in Figure 6 for iDoE batches, as well as for fixed condition experiments with $R < 0.4$. Experiment a ($R=0.45$) and q ($R=0.54$) were included to complete the set of fixed condition experiments presented in the designs of Figure 1. Qualitative inspection of the dynamic profiles of the training and validation batches confirms the respectable model performance for biomass and specific productivity. Biomass concentration profiles show a typical fed-batch fermentation trend in all batches: i.e. a positive but declining specific growth rate. Specific growth rates are affected by changes in process conditions (accounted for by the hybrid model) causing the batch to batch variations. Biomass concentrations are generally accurately predicted, with some minimal disagreements towards the end of some batches.

Generally, the productivity profiles show that cytoplasmic expressed product is quickly accumulated over the first 10 hours of induction at slightly different rates: i.e. early induction phase conditions are determining product formation. However, with the accumulation of product, the expression mechanism seems to be altered. According to Glick et al. [27], the metabolism of the host cell can change significantly during heterologous gene expression, often

a result of many physiological changes. In our case the specific productivity reaches a maximum or plateau around 0.6 to 0.8 OD⁻¹.

Specific product formation is well described by the model, except for under-predictions in experiments [c], [k] and III, which were discussed before. The agreement between measured and estimated productivity for these batches is acceptable during early induction phase, but model performance seems to worsen when the culture is presumably going through metabolic changes (the combination of high feed rate and low temperature encountered in these experiments, see Figure 1A, is not favorable and may lead to an overflow) and where product expression rate is quickly decreasing (Figure 6: III, [c] and [k]). One possibility is that these transitions are not properly described by the model, in this case resulting in under-prediction of productivity during the later stage of batch. The model may not take into account the influence uncontrolled process parameters. For example rates are considered dependent on X , P/X , T , pH and u_C , however concentrations of culture metabolites are not incorporated. They could be of particular importance in a transition region, where additional experimental data may benefit model performance.

A relatively high risk was observed for many experiments in the initial phase (see e.g. the risk profile of experiment [p] in Figure 4), which however does not seem to compromise the accuracy of the state predictions. The cutoff risk level ($R=0.4$), chosen based on the time averaged risk obtained for the validation set, is appropriate to ensure reliable prediction performance. The 4 iDoE experiments I, II, IV and V can mimic 12 different fixed conditions experiments (experiments [b] to [p]). They cover a space with five levels in temperature and

three levels in the feed rate. In addition, four levels in the induction biomass concentration (40, 50, 60 and 70 OD) are partially covered as well as a small pH range from pH 0.9 to 1.

Discussion of the findings

Process excitations such as step-changes, pulses, ramps or oscillations have traditionally been used by process control engineers to identify the response behavior of systems, referred to as system identification [11]. Yet, to the best of our knowledge, this concept has not been used in combination with design of experiments for upstream bioprocess development and optimization. We observed that 12 experiments executed at distinct conditions could successfully be described by a model that was developed upon four iDoE experiments.

However, one shortcoming observed was that the validity region did not include all experiments that were part of the fixed condition design space as would be expected. The region in which the model was expected to be valid ($R < 0.4$), did not include the conditions of experiment [a], which was part of the original investigated design space. Also iDoE experiment V, that spanned a large pH range, was not sufficient to augment the validity region to pH values lower than 0.9. The reason in both cases was that the experience measure domain in contrast to the design space comprises the biomass concentration and specific productivity, essential inputs of the first neural network. Due to its dual role as a model input and output it is difficult to include specific productivity as a factor in the design space, however the counter-rotating approach appears able to account for variations in the productivity in a large design space.

Despite the respectable results obtained in this feasibility study, it should be mentioned that dynamic variations in process conditions can result in transitions between intracellular (steady)

states [28, 29]. While these transitions can occur during constant process conditions, simultaneous excitations in pH, temperature and the feeding rate might be more likely to provoke state transitions, since these process variables are interconnected [30].

State transitions may result in different responses for the same process conditions. This behavior could have been a contributing factor in batches III, [c] and [k], where the productivity was under-predicted. To verify these mispredictions are due to state transitions, other extracellular metabolites such as acetate, pyruvate, lactate or succinate and/or the activity of certain enzymes [28, 29] should be measured. Batches conducted in a process condition space that includes a meta-stable transition zone will likely have cultures go through different cellular states. Static MVDA models, typically applied in process development, will likely not identify state transitions, only higher or lower biomass and productivity values between experiments.

The dynamic hybrid modeling framework applied in this study, has the potential to capture and describe state transitions, if sufficiently frequent measurements are taken during the process.

In light of the increasing application of online analytics [3, 7, 9] such as spectroscopic methods, frequent and detailed information about the process state will become available to support dynamic modeling. Knowledge about process dynamics is critical for advanced process control, such as closed-loop product quality control promoted in PAT and QbD.

Conclusions

Step-changes in temperature, pH and substrate feeding rate were performed in five *E. coli* fermentations to study whether intra-experiment variations of process conditions can be

applied to characterize the process operation space in a design of experiment manner: i.e. intensified design of experiments (iDoE). In our case, the operation space spanned by four experiments is equivalent to the process space characterized by a conventional design of experiments with three levels in the feeding rate, four levels in temperature plus two to three levels in the initial biomass concentration. Thus, it seems possible to characterize the process operation space with a minimal number of iDoE experiments. Key to this approach was a dynamic hybrid model that accurately estimated changes in biomass and product concentrations, specific productivity and base consumption. However, a region (at low temperature and high carbon substrate feeding rate) was encountered in which the accuracy of product concentration predictions deteriorated. It was hypothesized that the physiological state of the cell might have changed in this region. The more likely occurrence of state transitions for dynamic DoEs and the opportunity for their detection and modeling through the increasing use of online analytics has also been discussed. An experience measure was applied to determine a model validity domain, in which the model can be expected to predict process states accurately. Results of fixed condition fermentations performed within this domain were compared with model outputs to confirm this validity. However, the selected validity measure may to be too conservative for some process regions as the model provided accurate predictions for several experiments with higher risk.

Apart the significant reduction in the number of experiments needed to characterize an experimental region, it was argued that the combination of intra-experiment variation with dynamic modeling has the potential to improve process understanding, which is an essential component to reach the ultimate PAT goal of closed-loop product quality control.

For CQAs for which no process knowledge is available (such as product variants), the development of a hybrid model might be difficult. The analysis of the data, which is generated with the iDoE strategy requires a model that takes the complete time course of the process into account. It is therefore more difficult to develop static models that predict an end-point process property based on experiment wise fixed control degrees of freedom, such as Response Surface Models. A strategy for the automatic planning of the iDoE has not yet been developed, but will be essential for planning iDoEs involving multiple factors.

Competing interests

Jan-Martijn Hamelink is an employee of the GSK group of companies. The other authors declare that no competing interests exist.

Acknowledgements

The authors thank Chi Chung (Tim) Choi and Paul Rice for their collaboration in batch production and performing off-line sample analysis and Carl Hémond for providing the RP-HPLC assay and results. The authors thank Paul Rice also for critical reading of the manuscript. The author MVS thanks the Portuguese Foundation of Science and Technology for financial support (Ref.: SFRH/BPD/84573/2012).

References

- [1] FDA, Guidance for Industry -PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance. *Pharmaceutical CGMPs* 2004.
- [2] FDA, *United States Federal Food and Drug Administration* 2011.

- 443 [3] Rathore, A. S., Winkle, H., Quality by design for biopharmaceuticals. *Nat Biotech* 2009, 27,
444 26-34.
- 445 [4] Glassey, J., Gernaey, K. V., Clemens, C., Schulz, T. W., *et al.*, Process analytical technology
446 (PAT) for biopharmaceuticals. *Biotechnology Journal* 2011, 6, 369-377.
- 447 [5] Streefland, M., Martens, D. E., Beuvery, E. C., Wijffels, R. H., Process analytical technology
448 (PAT) tools for the cultivation step in biopharmaceutical production. *Engineering in Life Sciences*
449 2013, 13, 212-223.
- 450 [6] Ferreira, A. P., Tobyn, M., Multivariate analysis in the pharmaceutical industry: enabling
451 process understanding and improvement in the PAT and QbD era. *Pharmaceutical Development*
452 *and Technology* 2015, 20, 513-527.
- 453 [7] Mercier, S. M., Diepenbroek, B., Wijffels, R. H., Streefland, M., Multivariate PAT solutions for
454 biopharmaceutical cultivation: current progress and limitations. *Trends in Biotechnology* 2014,
455 32, 329-336.
- 456 [8] Mandenius, C.-F., Brundin, A., Bioprocess optimization using design-of-experiments
457 methodology. *Biotechnology Progress* 2008, 24, 1191-1203.
- 458 [9] Chew, W., Sharratt, P., Trends in process analytical technology. *Analytical Methods* 2010, 2,
459 1412-1438.
- 460 [10] Tai, M., Ly, A., Leung, I., Nayar, G., Efficient high-throughput biological process
461 characterization: Definitive screening design with the Ambr250 bioreactor system.
462 *Biotechnology Progress* 2015, 31, 1388-1395.
- 463 [11] Seborg, D. E., Mellichamp, D. A., Edgar, T. F., III, F. J. D., *Process Dynamics and Control*, John
464 Wiley & Sons, New York 2010.
- 465 [12] Åkesson, M., Hagander, P., Axelsson, J. P., Avoiding acetate accumulation in *Escherichia coli*
466 cultures using feedback control of glucose feeding. *Biotechnology and Bioengineering* 2001, 73,
467 223-230.
- 468 [13] Schaepe, S., Kuprijanov, A., Simutis, R., Lübbert, A., Avoiding overfeeding in high cell
469 density fed-batch cultures of *E. coli* during the production of heterologous proteins. *Journal of*
470 *Biotechnology* 2014, 192, Part A, 146-153.
- 471 [14] Dietzsch, C., Spadiut, O., Herwig, C., A dynamic method based on the specific substrate
472 uptake rate to set up a feeding strategy for *Pichia pastoris*. *Microbial Cell Factories* 2011, 10, 14.
- 473 [15] Montgomery, D. C., *Design and Analysis of Experiments*, John Wiley & Sons 2008.
- 474 [16] von Stosch, M., Oliveira, R., Peres, J., Feyer de Azevedo, S., Hybrid semi-parametric
475 modeling in process systems engineering: Past, present and future. *Computers & Chemical*
476 *Engineering* 2014, 60, 86-101.
- 477 [17] Psychogios, D. C., Ungar, L. H., A hybrid neural network-first principles approach to process
478 modeling. *AIChE Journal* 1992, 38, 1499-1511.
- 479 [18] Oliveira, R., Combining first principles modelling and artificial neural networks: a general
480 framework. *Comput. Chem. Eng.* 2004, 28, 755-766.
- 481 [19] von Stosch, M., Davy, S., Francois, K., Galvanauskas, V., *et al.*, Hybrid modeling for quality
482 by design and PAT-benefits and challenges of applications in biopharmaceutical industry.
483 *Biotechnology Journal* 2014, 9, 719-726.
- 484 [20] von Stosch, M., Hamelink, J.-M., Oliveira, R., Hybrid modeling as a QbD/PAT tool in Process
485 Development: An industrial *E. coli* case study. *Bioprocess and Biosystems Engineering* in press.

- [21] Gnoth, S., Jenzsch, M., Simutis, R., Lübbert, A., Product formation kinetics in genetically modified *E. coli* bacteria: inclusion body formation. *Bioprocess and Biosystems Engineering* 2007, 31, 41-46.
- [22] Gnoth, S., Simutis, R., Lübbert, A., Selective expression of the soluble product fraction in *Escherichia coli* cultures employed in recombinant protein production processes. *Applied Microbiology and Biotechnology* 2010, 87, 2047-2058.
- [23] Peres, J., Oliveira, R., de Azevedo, S. F., Bioprocess hybrid parametric/nonparametric modelling based on the concept of mixture of experts. *Biochem. Eng. J.* 2008, 39, 190-206.
- [24] Van Can, H. J. L., Te Braake, H. A. B., Dubbelman, S., Hellinga, C., *et al.*, Understanding and applying the extrapolation properties of serial gray-box models. *AIChE Journal* 1998, 44, 1071-1089.
- [25] Teixeira, A. P., Clemente, J. J., Cunha, A. E., Carrondo, M. J. T., Oliveira, R., Bioprocess iterative batch-to-batch optimization based on hybrid parametric/nonparametric models. *Biotechnol. Prog.* 2006, 22, 247-258.
- [26] Kahrs, O., Marquardt, W., The validity domain of hybrid models and its application in process optimization. *Chemical Engineering and Processing: Process Intensification* 2007, 46, 1054-1066.
- [27] Glick, B. R., Metabolic load and heterologous gene expression. *Biotechnology Advances* 1995, 13, 247-261.
- [28] Song, H.-S., Ramkrishna, D., Complex Nonlinear Behavior in Metabolic Processes: Global Bifurcation Analysis of *Escherichia coli* Growth on Multiple Substrates. *Processes* 2013, 1, 263.
- [29] Lei, F., Rotbøll, M., Jørgensen, S. B., A biochemically structured model for *Saccharomyces cerevisiae*. *Journal of Biotechnology* 2001, 88, 205-221.
- [30] Shimizu, K., Metabolic Regulation of a Bacterial Cell System with Emphasis on *Escherichia coli* Metabolism. *ISRN Biochemistry* 2013, 2013, 47.

Illustrations and figures

Figure 1 A: Two factor Doehlert design: five levels of temperature and 3 levels of carbon substrate feed rate and constant pH=1.0 and induction OD of 40: batches [a] to [g] ([g] was performed in duplicate designated by [g₁] and [g₂]). Intensified DoE: 4 fermentations, each fermentation covering 4 different conditions: batches I to IV. Set-points were changed four times over a 30 h induction period: at t=0, 4, 8 and 23 h. Batches [h] to [p] overlap with this design space. Batches indicated with a * were performed at lower pH and/or higher induction OD. B: One iDoE fermentation partially covering a 3-dimensional design space for temperature, substrate feed rate and pH: batch V. Set-points were changed three times over a 30 h induction period: at t=0, 6 and 24 h.

Figure 2: Hybrid model regression plots for biomass, specific and volumetric productivity and for accumulated base consumption. Training set (red star): batch II, IV and V; validation set (blue cross): batch I and test set (green x): batch III. Mathematical symbols as in the text.

Figure 3: Contours of membership function, $m(x_1(\tau), c_{n_c}, \sigma_{n_c})$ (grey lines), training data (red star), validation data (blue cross) and test data (green x) for combinations of neural network inputs.

Figure 4: Risk of model misprediction, $r(\tau) = 1 - m(x_1(\tau), c_{nc}, \sigma_{nc})$, as function of induction time for iDoE batches (training, validation and test data), as well as three experiments executed at different fixed process conditions.

542

543 Figure 5: Model performance for biomass and productivity as function of time averaged risk, R .
 544 Batch average Weighted Mean Squared Error (MSE) values are plotted. Upper row: biomass
 545 and accumulated base addition. Bottom row: specific and volumetric productivity. Red star -
 546 training data; blue cross – validation data and green x – test data.

Figure 6: Dynamic profiles for biomass (hybrid model - continuous black line, experimental data – black squares) and specific productivity (hybrid model - dashed dark-grey line, experimental data – dark grey circles). Experiments presented: iDoE batches I to V, fixed condition batches with a time average risk, $R < 0.4$ as well as experiment [a] ($R = 0.45$) and [q] ($R = 0.54$).

Tables and captions

Table 1: Description of process conditions of fermentations with varying intra-experimental conditions (iDoE batches) and fixed conditions. Experiments are sorted by increasing time averaged risk values, R . Designated batch numbers are as in the text.

Batch		Process Conditions				
	OD_{ind}	T	pH	u_c	Risk R	MSE
intensified DoE batches						
IV	43	-1.5 to 0.0	1.0	0.1 to 1.0	0.002	0.003
II	41	-1.0 to 0.5	1.0	1.0 to 1.9	0.003	0.010
V	41	-1.0 to 1.0	-1.0 to 1.0	-1.0 to 1.0	0.004	0.002
III	43	-1.5 to -0.5	1.0	0.1 to 1.9	0.34	0.43
I	41	-0.5 - 0.5	1.0	0.1 to 1.9	0.35	0.02
constant condition batches (Risk - $R < 0.4$)						
[f]	38	0.0	1.0	0.1	0.07	0.12
[h]	49	0.3	0.9	0.0	0.14	0.03
[p]	43	-1.0	1.0	1.8	0.14	0.06
[c]	40	-1.0	1.0	1.9	0.14	0.38
[e]	40	-1.0	1.0	0.1	0.15	0.05
[m]	44	-1.0	1.0	2.0	0.18	0.07
[o ₁]	40	-1.0	1.0	2.0	0.18	0.03
[o ₂]	39	-1.0	1.0	2.0	0.18	0.02
[o ₃]	39	-1.0	1.0	2.0	0.18	0.03
[o ₄]	39	-1.0	1.0	2.0	0.18	0.05
[j]	69	0.3	0.9	0.0	0.19	0.16
[n]	44	-1.0	1.0	0.0	0.21	0.09
[g ₂]	39	-0.5	1.0	1.0	0.22	0.08
[i]	62	-0.6	0.9	0.0	0.23	0.09
[g ₁]	38	-0.5	1.0	1.0	0.23	0.04
[b]	36	0.0	1.0	1.9	0.25	0.02
[d]	40	-1.5	1.0	1.0	0.28	0.11
[k]	40	-1.5	1.0	1.5	0.36	0.23
Batch		Process Conditions				
	OD_{ind}	T	pH	u_c	Risk R	MSE
constant condition batches (Risk - $R > 0.4$)						
	61	0.6	-0.1	-0.8	0.45	0.37
[a]	35	0.5	1.0	1.0	0.45	0.25
	51	0.9	0.0	0.0	0.48	0.53
	59	0.0	0.6	-0.8	0.50	0.27
	50	-0.9	0.0	0.0	0.52	0.10
[q]	38	0.0	0.0	0.0	0.54	0.06
	71	0.9	0.0	0.0	0.55	0.52
	51	-0.3	-0.9	0.0	0.57	0.16
	50	-0.3	-0.1	-0.8	0.58	0.03
	55	0.0	0.0	0.0	0.62	0.20
	60	0.6	-0.9	0.0	0.64	0.34
	62	0.0	0.0	0.0	0.65	0.12
	56	-0.6	0.1	0.8	0.67	0.35
	49	0.3	0.1	0.8	0.70	0.07
	75	-0.3	-0.1	-0.8	0.70	0.10
	59	0.0	-0.6	0.8	0.71	0.10
	70	-0.9	0.0	0.0	0.72	0.40
	61	-0.6	0.1	0.8	0.73	0.38
	78	0.0	0.0	0.0	0.76	0.14
	73	-0.3	-0.9	0.0	0.77	0.31
	74	0.3	0.1	0.8	0.79	0.13