**TECHNISCHE UNIVERSITÄT WIEN**

# Dissertation

# Data Science Workflows for Biopharmaceutical Manufacturing Process Validation Stage 1

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

Prof. Dr. Christoph Herwig

Institut 166/4

Institut für Verfahrenstechnik, Umwelttechnik und Technischen Biowissenschaften

eingereicht an der Technischen Universität Wien
Fakultät für Technische Chemie

von

Dipl.-Ing. Thomas Zahel
Matrikelnummer: 0075108
Holzstraße 19, 4020 Linz

_____           _____
Wien, am                                        eigenhändige Unterschrift

**Acknowledgements**

First of all, I would like to thank Professor Christoph Herwig for his advice and motivating encouragement throughout this thesis. Especially I deeply appreciate his way of giving guidance on the hand but letting me develop my own ideas and my own way of presenting those ideas on the other hand. Without this kind support neither this thesis nor my scientific and personal development would have been possible.

Special thanks to my colleagues at Exputec Patrick Sagmeister, Valentin Steinwandter, Stefan Hauer and Lukas Marschall. Without Patrick's imaginative ideas and visionary advice, without Valentin's extraordinary and enviable programming skills and without Stefan's patience and graphical skills this thesis would have been impossible. Especially, heartfelt thanks to Lukas Marschall for his thoughtful discussions and friendship during developing ideas for this thesis.

Big thanks to the entire Exputec team and Ulrich Tröller, who gave me the opportunity to discover industrial challenges during work that need to be solved scientifically.

Furthermore, I would like to thank Thomas Natschläger for his invaluable statistical discussions, clear thoughts and support.

Finally, big thanks to my family and friends, my parents Brigitte and Bert, and my grandparents, Paula and Gottfried, who provided me with place and time for me and my thesis to develop. That cordial support made it possible to endure that endeavor.

# 1 Content

3

# 2 Abstract

The biopharmaceutical market is innovative, well growing and delivering about 20% of all pharmaceutical product to patients. In order to consistently deliver high product quality the biopharmaceutical manufacturing process needs to be understood, controlled and effectively monitored. Those tasks are commonly addressed in manufacturing process validation, which is also requested from regulatory agencies due its importance in respect to patient risk. Especially the first step of achieving process knowledge by understanding and controlling potential sources of variance and risks is key to ensure successful routine manufacturing. Those activities are usually covered in process characterization studies (PCS) in industry.

Within this thesis, an advanced data science workflow for PCS is presented that points towards a holistic risk awareness and control strategy via knowledge obtained from single unit operations. Major novelties described in this thesis ensure on the one hand that information from single unit operations such as fermentation processes are accurately extracted. Moreover, novel statistical power analysis methods are presented to ensure that no critical information or process parameter on product quality has been overlooked. On the other hand an integrated process model has been introduced that facilitates to combine this knowledge from single unit operation by means of Monte Carlo simulation. The integrated process model was successfully applied on a real industrial process to derive holistic risk awareness and a holistic control strategy.

By applying this advanced workflow it is anticipated that variance in process output and product quality can be reduced and commensurately producers and patient risk is lowered.

# 3   Introduction

General aim in pharmaceutical manufacturing is to deliver drugs with consistent product quality in order to reduce patient risk of varying dose and quality. Therefore, variation and trends in product manufacturing must be understood, meaning that risk and impacting factors on quality need to be identified, controlled and ongoingly monitored [1]. Not exclusive driven by regulatory agencies, process validation has become the enabling task to achieve this goal [2]. The first guidance by the US Food and Drug Administration (FDA) appeared in 1987 defining process validation as *"Establishing documented evidence which provides a high degree of assurance that a specific process will consistently produce a product meeting its predetermined specifications and quality characteristics"* [3,4]. In 2011 an updated guidance document by the FDA appeared (entitled *Process Validation: General Principles and Practices*) as well as harmonized documents that aligned this process into three stages which are intended to be performed iteratively within a life-cycle approach [1,5,6]. Stage 1 defines the process design. It delivers the identification of interactions between process parameters (PPs) and critical quality attributes (CQAs) as well as the definition of a control strategy that aims to ensure high product quality. For definition of CQAs and PPs see ICH Q8 R2 guideline [5]. In stage 2, process performance qualification runs at manufacturing scale demonstrate the validity of the chosen design and capability to produce product quality. Stage 3 is where continued monitoring of defined critical quality attributes and parameters takes place to alert in cases of unexpected deviations from the state of control. Those stages are intended to be performed iteratively in a life-cycle approach. Whenever required, stage 1 or selected tasks from stage 1 can be repeated when unexpected variation has been observed or process changes have been introduced in stage 2 or 3.

Stage 1 is the place where process knowledge is formed, which is labor intensive and takes typically between 12 to 15 month [2,7]. Moreover, extended and scientifically sound activities at this stage are much more cost efficient than during later stages such as during investigation in counter action and preventive actions (CAPAs) [8]. An easy example is that investigation of an additional parameter and its interaction with other parameters in a DoE during stage 1 (e.g. going from 4 to 5 studied DoE factors) comes with an increase of only 2 experiments using for example definitive screening designs. Performing the same analysis at later stages (e.g. during a CAPA) would require to repeat the entire design with 12 experiments. Especially for biopharmaceutical products (recombinant proteins, virus as well as nucleic acids as defined elsewhere [9]) stage 1 is even more effort intensive compared small molecule products. This is due to:

- The process cannot be studied at manufacturing scale but small scale models need to be employed and checked for being representative [10]

- characterization of the product is analytically complex and associated with higher variance
- reliable analytical measurements of product quality are hard to obtain at early (upstream) unit operations
- host cells for fermentation are black to grey boxes with complex metabolic activities
- The manufacturing process is composed of multiple unit operations which highly increases the number of potential sources of variance.

Currently, the worldwide pharmaceutical market is about $1 trillion with 40,000 ongoing (or recently reported) candidate products in clinical trials [11]. About 40% of the candidate products are likely biopharmaceuticals. Assuming that each newly developed product needs to undergo one or the other way of process validation, it can be expected that a huge effort and money is spent in those activities. Due to the complexity and the expected industrial need this work will focus on biopharmaceutical process validation stage 1.

Biopharmaceutical process validation stage 1 activities are frequently called "process characterization studies" in biopharmaceutical industry, although this is not a regulatory term. They try to accomplish following goals of stage 1 process validation [12]:

- Understanding of the role of each unit operation to deliver quality product
- Identification of the impact of PPs onto CQAs
- Setting of a control strategy for critical process parameters (CPPs)
- Setting acceptance criteria for in process control even on intermediate unit operations

A workflow for that endeavor is present in industry standard operating procedures (SOPs) and literature, which is divided into (i) data mining (ii) risk assessment (iii) scale down qualification (iv) characterization of process performance and (v) experimental studies to reveal dependencies of potential critical process parameters (pCPPs) onto CQAs (Figure 1) [7,12].
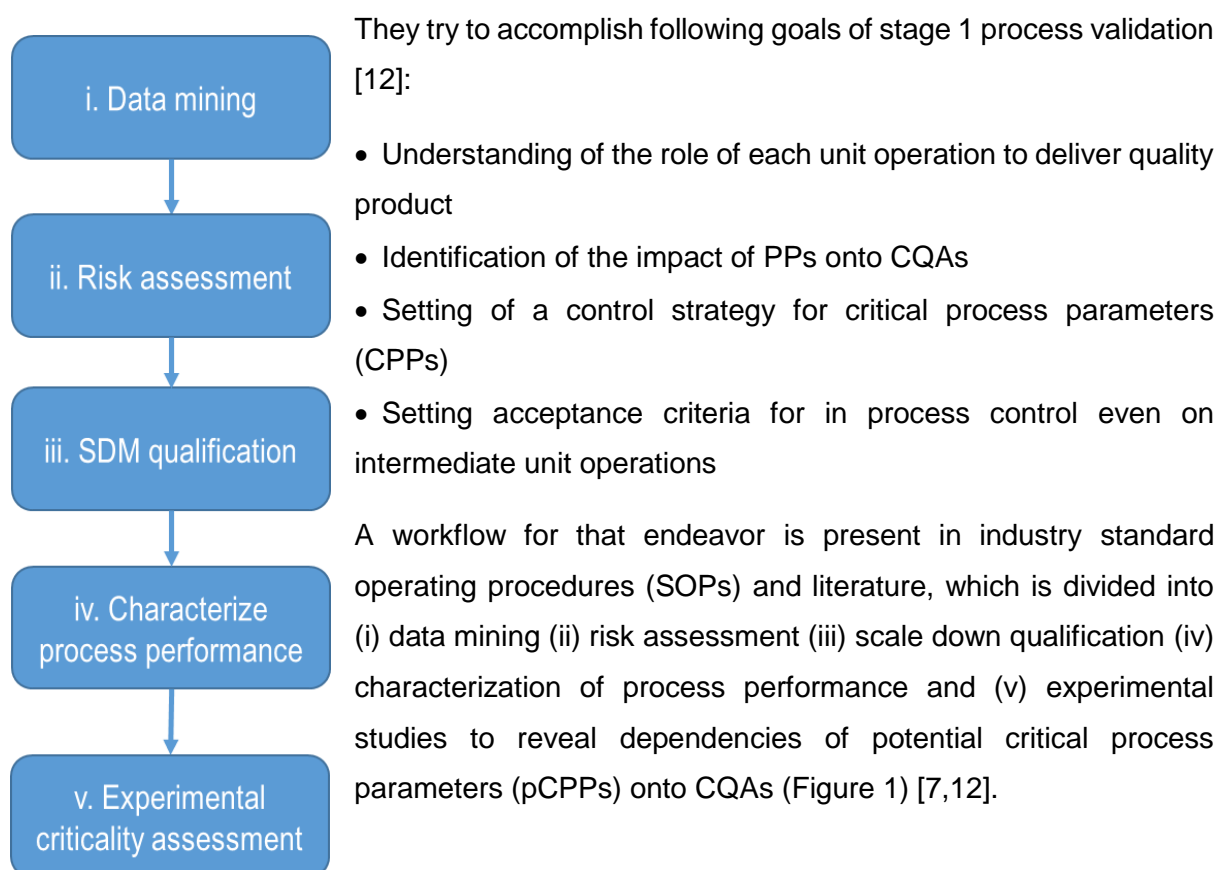


*Figure 1: Typical process characterization workflow as suggested by Seely and Seely [7]*

## 3.1  Data mining and risk assessment

Goal of data mining and risk assessment is to identify potential impacting process parameters onto CQAs using indications from historical data (development or manufacturing data) and knowledge of process experts. A multitude of risk assessment tools are commonly applied in industry such as, hazard analysis and critical control points (HACCP) [13] or failure mode, effects and criticality analysis (FMECA) [14,15]. Those tools rate the criticality of the parameters according to the occurrence, harm and detectability of the potential risk. Many of those approaches are mature as they are frequently applied in industry, however, a direct link to the development data in the final risk assessment document is hardly established. Linking risk assessment directly to data or additional evidence in a document structure might reduce the bias of individual persons involved in the risk assessment group.

## 3.2  Scale down model (SDM) qualification

Scale down models are commonly used to investigate effects of the manufacturing scale at small scale. This is necessary since manufacturing scale experiments are not feasible for biopharmaceutical processes. Some unit operations cannot be scaled down easily and must be investigated in pilot scale (e.g. separation of biomass). For each unit operation scale down principles exist that describe parameters that need to be kept constant during scale down (e.g. bed height, height equivalent of theoretic plates and peak asymmetry for chromatography, specific power input for microbial fermentations or trans-membrane pressure for ultrafiltration/diafiltration) [16,17]. Although those essential parameters are kept constant, scale difference might exist due to inhomogenieties or differences in equipment [18,19]. The potential impact of those effects on product quality needs to be examined. This is done by comparing performance parameters between small and manufacturing scale, which are descriptive for the role of each unit operation for obtaining quality specifications. This process of selecting suitable responses for comparison is performed separately during characterization of process performance, which is sometimes proposed after SDM qualification (see Figure 1). However, this ordering might be rearranged since response selection, according to process and impurity clearance performance, should be performed before SDM qualification in order to have descriptive responses in place for scale comparison.

Only a valid SDM is predictive for the effects of PPs onto CQAs at large scale. This is also reflected in the ICH Q11 guideline [20]: *"The contribution of data from small-scale studies to the overall validation package will depend upon demonstration that the small-scale model is an appropriate representation of the proposed commercial-scale. Data should be provided demonstrating that the model is scalable and representative of the proposed commercial process."* Therefore, a truly representative SDM shows the same effects that would occur in manufacturing scale [10]. Different possibilities of effects and offsets in manufacturing and small scale are shown in Figure 2. Since experimentally investigating effects in manufacturing

scale is not feasible, it is common practice to compare all responses between the scales at target operating conditions. If there is no difference at target conditions (cases A and B in Figure 2) we assume from a heuristic point of view that it more likely to have similar effects (case A). This is due to the fact that the probability of finding the intersection of diverging effects at an arbitrarily chosen point in the design space (here target conditions) is low. On the other hand if there is a practical relevant difference between the scales as shown in cases C and D of  Figure 2, both cases seem possible and the risk associated with finding an non-predictive model of case D needs to be analyzed, estimated and mitigated. A possible solution to mitigate the risk is to identify root causes for the difference of scales at target operating conditions and estimate the likelihood that this impacts on differences in effects, too.
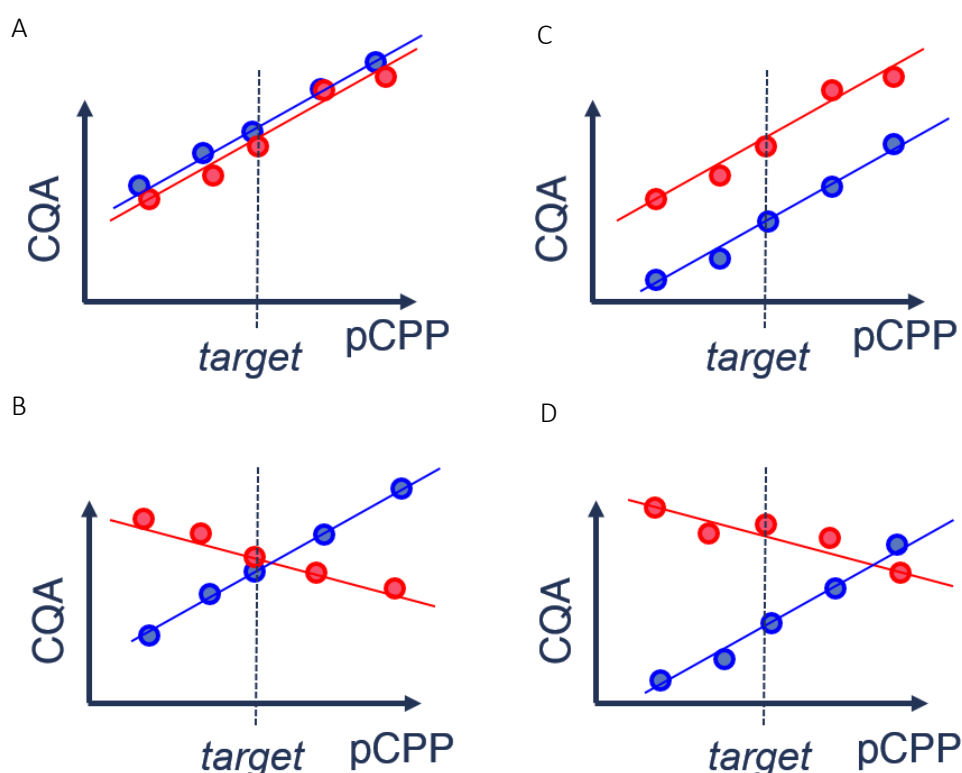


*Figure 2: Taken with kind permission of Exputec GmbH [21]. Predictability of scale down model (blue) to large scale (red). A: no offset good predictability. B: no offset but different effects from large and SDM leads to bad predictability. C: shows good predictability although a constant offset exists at target. D: shows bad predictability of scale down model where a similar offset at target exists.*

For qualitative performance indicators such as turbidity, SDS gels, curvature of preparative chromatograms a side-by-side comparison is usually conducted. For quantitative parameters statistical approaches are used to assess whether small and manufacturing scale can be regarded practically equivalent at target conditions [12]. A variety of statistical methods exist such as testing if all small scale runs are within minimum to maximum (min/max) of manufacturing runs, two sample t-tests of difference in means [12], intersection-union tests [22], two one sided t-tests (TOST) [23] or even multivariate approaches [24,25]. However, their limitations are hardly fully understood by users. For example, when testing if all small scale

runs are within min/max or 3 standard deviations of the large scale runs, we reward ourselves for large scattering in the manufacturing runs and potentially overlook practical relevant differences of the SDM. When using two sample t-test of difference in means we are not able to distinguish between practical relevant and statistical significant differences. This leads to false alarms by detection non-relevant differences, and even more critical, to overlooking of practical relevant differences. Therefore, TOST has been established to detect practical relevant differences in the mean performance of the scales by comparing the confidence interval of the mean difference between small and manufacturing scale to a pre-defined equivalence acceptance criteria (EAC) [26]. The practical inconvenience is often caused by the hurdle of defining EACs, which should not be established using the data of either small nor manufacturing scale, but should be established using prior knowledge of a practical relevant difference. Recently also equivalence testing approaches (also entitled comparability analysis) have been reviewed by regulatory agencies [27–30]. However, also those guidance documents lack in describing which approaches can be employed to define practical relevant EACs for comparing manufacturing processes, which is the pivotal question in equivalence testing.

## 3.3  Characterization of process performance

Each unit operations or multiple unit operations are intended to maintain some sections of the quality target product profile (QTPP, as defined in ICH Q8 [5]) [2]. Characterization of process performance therefore aims to identify which unit operation is responsible for reaching specification limits for specific CQAs. Furthermore, it is necessary for each unit operation to identify key quality performance indicators that are used during experimental studies and SDM qualification as responses [7,12].

For downstream unit operations typically a majority of the CQAs can be measured analytically before and after the unit operation, therefore, yields and clearance factors of impurities are compared as key performance indicators between the scales. However, in upstream unit operations (i.e. fermentation), CQAs can hardly be directly assessed. FDA clearly states in the process validation guideline, that special attention should be paid to operational limits and in process controls *"when the product attribute is not readily measurable due to limitations of sampling or detectability"* [6]. Herein, special attention needs to be paid to the upstream unit operations. The physiological state of host cells triggers metabolic activity that impact on product quality (e.g. glycosylation, monomer-dimer formation, disulfide formation) [31,32]. Therefore physiological information about the production organism is a suitable indirect measure for product quality. Physiological information about the cells is most effectively estimated by cell specific turnover rates. Complexity is increased since cell specific turnover rates are more complex to estimate than overall yields of a downstream unit operation and,

moreover, are dynamically changing over time. This is especially pronounced for the dominant process mode in industry, the fed-batch fermentation. Although specific rate calculation is based on derivatives from mass balances and can be found in many text books, their accuracy might vary in respect to the accuracy of the original data, which they are derived from (concentration of chemical species and biomass) [33]. Original concentration data might be measured analytically by taking samples out of the reactor ("offline data") or estimated indirectly from online signals via software sensors [34–37]. In both cases measurements might be error prone and it was unclear how this propagates to error in specific rates. Therefore, two previously un-addressed gaps have been identified and targeted in this work:

- How can specific rates from offline measurements (dry cell weight, viable cell density) be calculated with constant signal to noise ratio? This reduces the risk that decisions are made upon data with little accuracy [38].
- How can this be achieved and for software sensors that use off-gas measurements to non-invasively estimate the biomass? [39]

## 3.4 Experimental identification of CPPs

Goal of experimental studies is to identify which potential critical process parameters (pCPPs) have practical relevant impact on product quality. Subsequently, a sufficient control strategy must be developed for those parameters and thereby reduce the risk of process failures and risk to the patient.

This is also defined in ICH Q8 [5] where it is stated that *"The list [of pCPPs] can be refined further through experimentation to determine the significance of individual variables and potential interactions."* A variety of experimental strategies are used to explore the impact of PPs onto CQAs. One possibility is to use screening designs to select active main effects, followed by a refinement study using response surface designs to estimate interaction and quadratic effects [7]. On the other hand it is also possible to use D-optimal designs or special forms such as definitive screening designs to study main effects, interaction and quadratic effects within a single design [40–42]. The latter one has the benefit that it is possible to study all two factor interaction effects and quadratic effects of potential critical process parameters at the price of little cofounding of two factor interactions with their respective quadratic effects.

To select CPPs, which have practical significant impact onto the CQAs, from the set of significantly impacting PPs different rules (Z-score, 20% rule) can be applied [43]. However, if a parameter is not statistically significant it does not necessarily mean that it has no impact. This seems contra-intuitive at the first glance but on a higher level we can say "if we do not see it, it does not mean that it is not there". Especially when analytical variance is hindering to detect effects it is likely that we overlook a critical effect. This has not been payed attention to in the past, however, it bears a huge potential for unknown risk to the process. Statistical power

10

of a test or a DoE is the chance to detect a significant effect if it is truly there, and it is therefore a suitable measure to estimate the risk of overlooking a critical PP. Power is positively influenced by the number of experiments, the significance level of the test and the signal to noise ratio of the experimental measurements. Statistical power analysis conducted before starting the experiments ("*a priori*" power analysis) is a mature technique to investigate the expected power in advance and usually adapt low powered settings by increasing sample size [44,45]. Rarely the assumptions of the expected signal to noise ratio during a priori power estimation are checked after the experiments have been conducted (retrospective power analysis). This is also due to the fact that a suitable statistical method to estimate power for individual parameters retrospectively in a multivariate model has been missing [46,47].

Therefore, it was aim of this work to develop an algorithm that is able to identify potentially overlooked effects and answer the following question with terms from project risk management [48]: How can we convert "unknown unknown" CPPs to "known unknown" CPPs?

## 3.5  Holistic risk assessment

Risk and criticality of single PPs and material attributes is usually assessed in a unit operation wise manner. Since biopharmaceutical manufacturing processes are composed of multiple unit operations it is necessary to holistically identify and analyze the risk of a failure and potentially adapt the control strategy. This is usually a task addressed by dynamic and steady state flowsheet modeling in process system engineering [49–51]. Here first principles, mechanistic and empiric equations including a set of parameters describe the interaction between PPs and in-process controls or CQAs for each unit operation. Those models are typically applied to unit operations where the mechanistic understanding is high, i.e. where impact of PPs onto CQAs can be written by ordinary differential equations (ODEs) or differential algebraic equations (DAEs). Those unit operations are usually final steps of oral solid dosage form production (crystallization, filtration, drying, blending or tablet pressing) [52,53]. However, typically for a biopharmaceutical manufacturing process mechanistic knowledge of the impact of PPs onto CQAs is rarely available for unit operation prior to drug substance (e.g. fermentation) or parameters of the mechanistic models are hard to identify. Furthermore, in process system engineering variation in the process input is rarely addressed and therefore a probabilistic estimation of variance in process output cannot be estimated. Herein, we want to present a possibility how to concatenate statistic (black box) models of each unit operation and estimate the overall process capability at normal operating variance of the process parameters using an integrated process model (IPM).

An additional aim of process characterization work is also setting of acceptance limits even at intermediate unit operations. Those limits, defined on in-process-controls and intermediate CQA measurements, must not be exceeded in order to reach final drug substance or product

specifications. One approach is to use estimates of the scattering of historical manufacturing data (e.g. 3 standard deviations or tolerance intervals) [54,55]. However, deriving acceptance limits from existing data does not take potential process capability into account. One the one hand we might overestimate process capability, especially when using tolerance intervals with low sample sizes (e.g. only 3 manufacturing runs). In this case the acceptance limits are too wide and it is likely not to reach drug substance specifications. On the other hand, a limited set of manufacturing runs do not represent a homogeneous sample of the entire design space since they are run at target conditions with little variance in PPs. In this case the acceptance limits are too narrow and we obtain many false positive deviations leading to investigations. Therefore, acceptance limits should be established taking holistic process capability into account by using all available knowledge obtained during historical manufacturing but also during process characterization studies. However, a quantitative methodology has been missing so far. Therefore, we want to show how the IPM can be used to fill this gap.

## 3.6   Goals of an advanced data science workflow for process validation stage 1

Despite the effort spent on defining workflows and SOPs for biopharmaceutical process validation stage 1, many validated processes still show high variance leading to deviations, investigation, trouble solving and patient risk [2,7]. Improvements to existing strategies are necessary to ensure that risk estimation is scientifically sound and on a quantitative level. Therefore, it is goal of this work to enrich state-of-the-art workflows in biopharmaceutical process validation stage 1 with data science tools that aid in understanding process variance, holistically estimating risk to the product and ultimately increasing process robustness.

I want to achieve this by projection of existing and development of new statistical methods that can be incorporated in existing process validation activities. Improvements to existing process validation stage 1 workflow aim to address following gaps which have been identified above:

- Responses selection: How to calculate physiological cell information with consistently high accuracy for biopharmaceutical fermentation processes and thereby lower risk in decision making?  (section 4.3)
- Experimental evaluation: How to identify unknown unknown risks and make them known unknowns? How to place a control strategy on identified known unknown critical process parameters (CPPs)? (section 4.5)
- Holistic process control and risk assessment: How can we estimate the risk of individual parameters or unit operations holistically in the process? (section 4.6)

Those questions have not been addressed from a statistical point of view and have not been considered to be applied for process validation. The individual improvements to those questions, their novelty and impact will be contrasted to existing approaches in section 4. By

12

mutually applying those novel methods, we anticipate that process robustness is increased, patient risk is lowered and manufacturing costs can be reduced.

Moreover, anticipated solutions to additionally identified gaps in equivalence testing for SDM qualification, will be presented in the outlook of this work (section 7).

# 4 Advanced workflow for biopharmaceutical manufacturing process validation stage 1

Figure 3 shows the proposed workflow for biopharmaceutical manufacturing process validation stage 1 including 6 sub-tasks with their respective connection of outputs of the previous task to inputs of the next task. Blue boxes indicate tasks, tools or deliverables that are currently applied in biopharmaceutical industry [7]. Green font colors in green boxes indicate areas to which this contribution will suggest improvements. At the end of this workflow after task 5 and 6 the final aim of stage 1 process validation, setting a control strategy, is achieved. Although Figure 3 depicts a series of discrete, numbered events, depending on the stage of the product life cycle, some steps might be skipped. E.g. in case unexpected variance was encountered during routine manufacturing in stage 3, the scale down model (SDM) might be still valid to perform representative experiments. Therefore, SDM qualification does not need to be performed again. However, the workflow is intended to show stage 1 process validation activities as performed for the first time for a new product.

In the following sections I want to review all single tasks of the proposed workflow as shown in Figure 3 together with their inputs and outputs. Moreover, I want to highlight how manuscripts of this work (section 5) can be used as an enabler for this workflow.

13

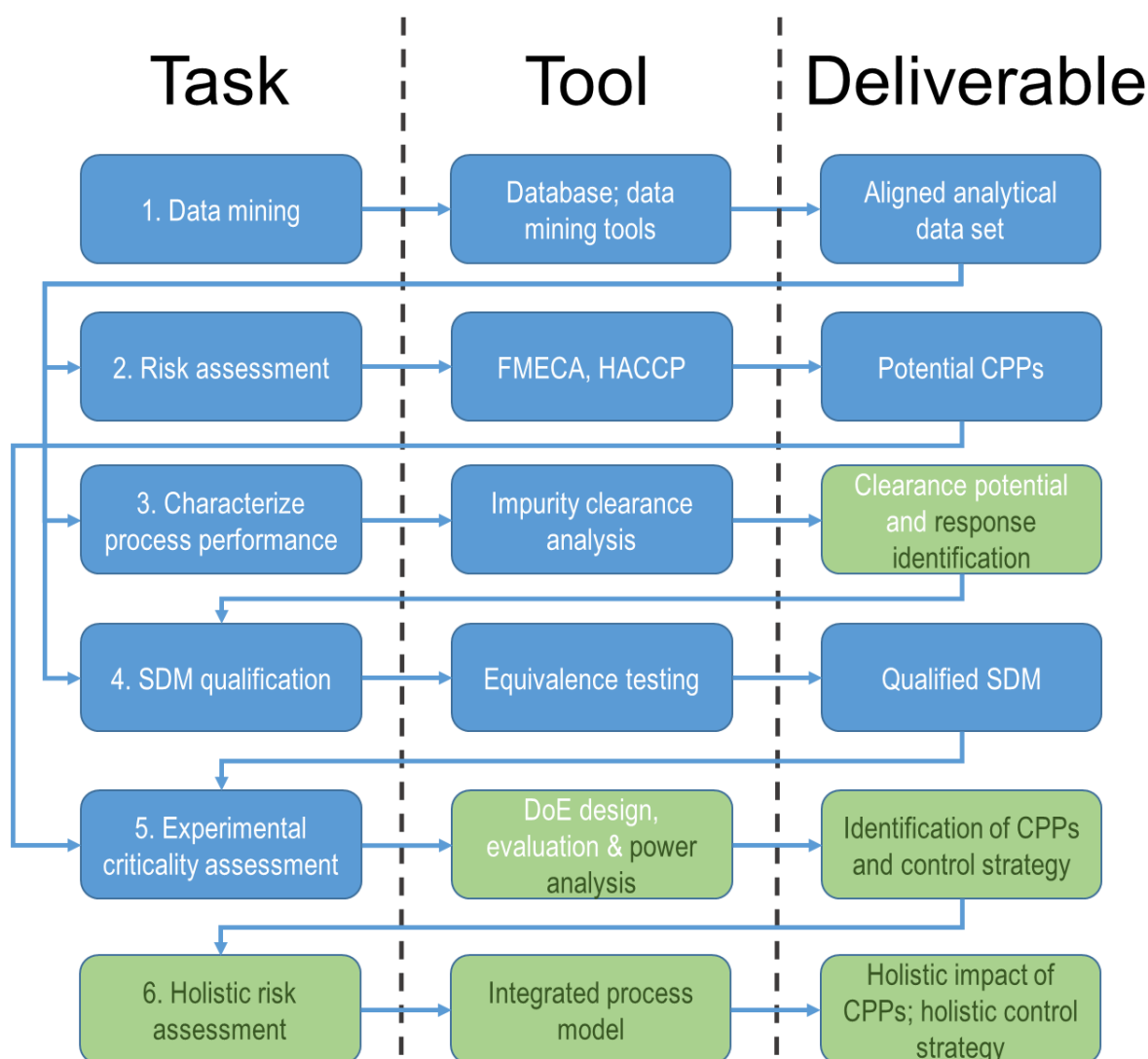| Task | Tool | Deliverable |
|------|------|-------------|
| 1. Data mining | Database; data mining tools | Aligned analytical data set |
| 2. Risk assessment | FMECA, HACCP | Potential CPPs |
| 3. Characterize process performance | Impurity clearance analysis | Clearance potential and response identification |
| 4. SDM qualification | Equivalence testing | Qualified SDM |
| 5. Experimental criticality assessment | DoE design, evaluation & power analysis | Identification of CPPs and control strategy |
| 6. Holistic risk assessment | Integrated process model | Holistic impact of CPPs; holistic control strategy |

*Figure 3: Workflow for biopharmaceutical process validation stage 1. Blue boxes indicate areas that are applied as a standard approach in industry (compare Figure 1). Improvements are presented to tasks, tools and deliverables indicated by dark green font color in green boxes. Arrows indicate connections of inputs and outputs for of individual steps.*

## *4.1* Task 1: Data Mining

Input:

- Development data
- Manufacturing data of product or similar product in case of a platform process

Output:

- Aligned analytical data set, which is input for task 2, 3 and 4

Data Mining is one of the first steps performed to gather development, pilot and large scale data. Input to this step are more (manufacturing data) or less (development data) structured data sources. Data types vary from complex time series data of fermentation and analytical and preparative chromatograms up to single valued quality data. Complexity of time series

14

data is increased due to shifts between single runs, e.g. in inoculation start for fermentation data or elution start for chromatograms. Therefore, databases and data mining tools are used frequently to align data in a single structured analytical data set. Requirements to that data set are that sparsity is reduced to a minimum, especially for CQAs, potential critical PPs (pCPPs). It is important to cover as many runs connected to the product or similar product in case of a platform process. This aligned data set facilitates tasks in process performance characterization and risk assessment. Moreover, it is intended to be extended when new data is recorded from scale down model qualification or design of experiments (DoEs) to be compared to existing data.

## 4.2   Task 2: Risk assessment

Input:

- Aligned analytical data from task 1

Output:

- Potential CPPs that will be studied in experimental assessment (task 5)

Manufacturing risk assessment is a crucial step to identify potential CPPs (pCPPs) using priori scientific knowledge about the process and the equipment. The definition of a CPP is given in the ICH Q8: "*A process parameter whose variability has an impact on a critical quality attribute and therefore should be monitored or controlled to ensure the process produces the desired quality*" [5]. General tasks in a risk management process to identify pCPPs are (i) risk assessment, (ii) risk control and (iii) risk review, as described in detail in ICH Q9 [1]. Risk control and risk review are addressed during the establishment of a control strategy and during stage 2 and 3 of the process validation, respectively. The first step, risk assessment in manufacturing, is usually carried out using tools such as Failure Mode Effect and Criticality Analysis (FMECA) or Hazard Analysis and Critical Control Points (HACCP). Due to the lack of extensive data, commonly evaluation of severity, occurrence and detectability of the harm is done on a qualitative basis using priori knowledge of process experts. Practically, this is carried out in team meetings which has the drawback that single persons can act as influencers on the overall team-opinion and thereby bias the output. Therefore, also the ICH Q9 guideline suggests to assemble background information. For example, occurrence of a failure mode (e.g. deviation of a process parameter beyond operating ranges) might be estimated from historical facility data (e.g. variance of a controlled process parameter). Using data driven risk assessment ensures that assignment of risk priority numbers (RPNs) to process parameters is less subjective and thereby the chance of overlooking a CPP is reduced. This is in line with the principle of risk management that "*the evaluation of risk to quality should be based on*

*scientific knowledge and ultimately link to the protection of the patient*", as defined in the ICH Q9 guideline.

The result of risk assessment is a set of pCPPs that might be studied in experimental criticality assessment. Methods for selecting this set range from taking all pCPPs that exceed a pre-defined RPN cutoff value or just applying a Pareto principle by taking the 20% most highly ranked pCPPs [7].

## 4.3 Task 3: Characterization of process performance

Input:

- Aligned analytical data set from task 1

Output:

- Responses that will be studied during SDM qualification (task 4) and as a function pCPPs during experimental criticality assessment (task 5)

Within impurity clearance reports the process performance in respect to product yield and impurity formation and clearance of individual unit operations is characterized using the established analytical data set. If a unit operation is responsible for the clearance or producing of a specific CQA, the performance in respect to this CQA should be taken as a response for experimental assessment during SDM qualification and DoE studies. In the easiest way all CQAs will be defined as response for each unit operation, however, due to analytical limitation it will not possible to measure all of them at load and pool of each unit operation. Moreover, if placing a CQA as a response for a unit operation which is not responsible for its purification or clearance, it might only raise false positive alarms of statistical non-similarity during SDM qualification or critical effects during experimental criticality assessment without practical meaning.

As described in the introduction, special attention should be paid to estimate the performance and physiological state of the fermentation of biopharmaceutical processes. This is the place where impurities are formed but where least of them can be measured analytically. Although physiological turnover rates form a good basis for monitoring metabolic differences that are connected to differences in product quality, they might be superimposed by large fraction of noise. Therefore, I want to present here two methods that ensure that calculated specific turnover rates from offline measured samples as well as indirect softsensor estimation of specific rates contain constantly high signal to noise ratios [38,39].

Using those methods it is possible to achieve risk awareness about the error of specific rates which are compared between small and manufacturing scale in task 4 or even can act as responses in experimental criticality assessment (task 5). Moreover, the associated error with

specific rates, which can be derived using the presented methods, can be used as intra-batch variability and can be compared to inter-batch variability of the runs. This is an important measure how likely we can observe a true difference between scales in task 4, or detect a significant effect in DoE studies in task 5.

## 4.4 Task 4: SDM qualification

Input:

- Aligned analytical data from task 1 to provide data from manufacturing scale, identified responses from task 3

Output:

- Identification of similarities and potential differences between SDM and manufacturing scale
- Sufficient evidence to state the SDM representative to the manufacturing scale

Since manufacturing scale experiments are costly to explore dependency of process parameters onto CQAs, it is common practice to use small scale experiments instead. The process of comparing the down scaled process to the manufacturing scale is called scale down model (SDM) qualification. Comparison is conducted on process performance based responses, which have been identified in task 3 of the proposed workflow (Figure 3). Therefore, in contrast to existing workflows, it is suggested to perform characterization of process performance before SDM qualification. Additional improvements to statistical equivalence testing for biopharmaceutical SDM qualification will be discussed in the outlook of this work in section 7.

## 4.5 Task 5: Experimental criticality assessment

Input:

- Aligned analytical data set from task 1, identified responses from task 2 and pCPPs to be studied identified by task 4

Output:

- Classification of pCPPs into CPPs and practically non-critical PPs
- Set a control strategy for PPs, potential tightening of tentative control strategy for CPPs

After a SDM model can be regarded representative for the manufacturing scale either by stating equivalence of all key performance parameters or identification of known root causes for the difference in scale, experiments can be conducted at small scale to assess the criticality of pCPPs. This process is carried out for each unit operation separately and will be in the

following entitled criticality assessment. Thereby, usually DoE studies are performed to reveal the impact of main effect, interaction and quadratic effects of pCPPs, defined in task 2 (risk assessment), onto process performance based responses, defined in task 3 (characterize process performance).

From experimental data, statistical regression models ($CQA = f(pCPPs)$) are built using stepwise regression or Pareto charts to select significantly impacting pCPPs (those which are not likely to have no effect) to enter the model. Currently, those parameters which do not enter the model are stated as non-critical, without having performed further analysis. This is unknown risk to the process since the chance of overlooking this PP as a CPP has not been estimated.

We developed a retrospective power analysis method that uses permutation tests to estimate the risk of having overlooked a CPP, which is explained in detail in one manuscript of this work [40]. This methodology gives a possibility to convert "unknown unknown" CPPs to "known unknown" CPPs [48]. Moreover, using the presented algorithm it has been successfully shown on a real manufacturing process how to impose a suitable control strategy on "known unknown" CPPs and thereby mitigate the risk of future process variance [40]. Therefore, we understand the definition of a CPP by ICH Q8 in a more probabilistic way, transforming the original statement "*parameter whose variability has an impact on a critical quality attribute*" into "*parameter whose variability is likely to have impact on a critical quality attribute*". If we do not assess this chance we potentially underestimate the risk of our process to the patient and have to pay a higher price during trouble shootings at later stages of process validation and commercial production when unexpected variation in product quality is encountered. This is then possibly due to an insufficient control of the "unknown unknown" CPPs. Moreover, the 2011 FDA process validation guideline states that criticality should be "continuum rather than a binary state" [6]. This can be interpreted in two ways. On the one hand, there is a smooth transition of non-critical to CPPs and on the other hand criticality of a parameter might be assessed multiple times during the lifecycle and is not fixed. During iterative cycles of stage 1, 2 and 3 unknown criticality might become visible or CPPs turn out to be less critical than initially assumed.

As an output of this subtask the critical PPs of each unit operation are identified and a suitable control strategy is placed on them in order to ensure that product quality can be continuously delivered. In future the presented power analysis can also be used to establish design spaces in quality by design (QbD) approaches [10,56,57].

## 4.6 Task 6: Integrated process model

Input:

- Aligned analytical data from task 1, statistical regression models from DoEs (task 5)

Output:

- Identification of holistic criticality of PPs and potential tightening of the control strategy
- Assessing overall process robustness
- Holistic and process capability based acceptance criteria for intermediate unit operations

In one contribution we employed a novel method that uses Monte Carlo simulation as a connector between statistical DoE models from task 5 of single unit operations [58]. Discrete combinations of possible PP combinations of multiple unit operations are propagated through the entire process and risk of out of specification events at batch release is estimated. The IPM is extendable to use mechanistic instead of statistical models, too. The approach assumes that no critical effect has been overlooked to be included in the statistical models, which can be ensured by applying the novel retrospective power analysis as presented in task 5. Otherwise, process robustness of the IPM might be overestimated since unknown risk has not been taken into account. Although Monte Carlo simulations are widely used their application in process validation stage 1 to assess process robustness and overall parameter criticality has not been reported before to the best of my knowledge [59,60]. The IPM has successfully been applied during industrial process validation to predict representative out of specification (OOS) probabilities under normal manufacturing variance of PPs. If OOS probabilities are low, this can be regarded as evidence that a process will likely be capable of consistently delivering quality product also in future. If OOS probabilities are high, control strategy can be tightened for those holistically impacting CPPs. Therefore, by application of the IPM it will be possible to lower the OOS probabilities most efficiently, offering benefits to manufacturer and patients. Additionally, it is has been demonstrated, using parameter sensitivity analysis, how to assess the holistic criticality and impact of single PPs onto product quality within a mutual interplay of the entire process.

Moreover, within the IPM framework it is possible to calculate performance based acceptance criteria for intermediate unit operations, which are required to monitor impurity and product content of future batches at intermediate process steps. Operating within those limits ensures reaching specification at batch release. Those acceptance criteria can be obtained via parameter sensitivity analysis of the IPM by varying CQA concentrations at the load of intermediate unit operations.

# 5  Manuscripts

## 5.1  Author contributions

| Manuscript | Contribution of Thomas Zahel (TZA) |
|---|---|
| Physiological rate calculation [38] | TZA designed the adaptive window rate calculation algorithm and the generic workflow for rate calculation for biopharmaceutical online and offline fermentation signals. |
| Softsensor error propagation [39] | TZA developed and implemented the error propagation procedure within the softsensor framework as well as the design of the generic workflow to identify tolerable measurement accuracy to deliver predefined softsensor accuracy. |
| Criticality assessment [40] | TZA developed the retrospective power analysis and criticality assessment workflow and wrote the manuscript. |
| Integrated process model [58] | TZA designed the IPM and the Monte Carlo simulation and wrote of the manuscript. |

## 5.2  Manuscripts

# Accurate Information from Fermentation Processes - Optimal Rate Calculation by Dynamic Window Adaptation

Authors

Thomas Zahel[1], Patrick Sagmeister[1], Szymon Suchocki[1] and Christoph Herwig[2, 3]

[1] Dipl.-Ing. Thomas Zahel, Dr.techn. Dipl.-Ing. Patrick Sagmeister, B.Sc. Szymon Suchocki, Exputec GmbH, Pfeilgasse 32, Vienna, Austria

[2] Univ.Prof. Dr.techn. Dipl.-Ing. Christoph Herwig, Institute of Chemical Engineering, Research Area Biochemical Engineering, Vienna University of Technology, Gumpendorferstrasse 1a, Vienna, Austria

[3] Univ.Prof. Dr.techn. Dipl.-Ing. Christoph Herwig, CD Laboratory on Mechanistic and Physiological Methods for Improved Bioprocesses, Vienna University of Technology, Gumpendorferstrasse 1a, Vienna, Austria

Correspondence concerning this article should be addressed to C. Herwig at christoph.herwig@tuwien.ac.at

Abstract

The calculation of metabolic turnover rates is essential for the scalable design, analysis and control of bioprocesses. Here, we present a novel rate calculation algorithm based on the dynamic adaptation of window sizes in order to deliver robust and precise rates with uniform signal-to-noise ratios. Moreover, we present a model-based generic algorithm for deriving optimal rate calculation workflows. The generic algorithms delivered more precise and accurate rates for on- and offline signals, which was demonstrated for both in silico- and real batch and fed-batch fermentation process data. The presented algorithms will strongly support bioprocess development and control as enabling tools for multivariate data analysis, mechanistic modelling and dynamic experimentation.

Keywords: bioprocess, bioprocess analysis, *Escherichia coli*, *Pichia Pastoris*, metabolic turnover rates

## 1. Introduction

Bioprocesses are key drivers for innovation in the pharmaceutical industry, whereby global sales of drugs produced in bioprocesses exceeded US $100 billion in 2010 [1]. Furthermore, industrial bioprocesses are considered key processes for achieving a sustainable bioeconomy [2]. In order to stay competitive, a constant increase in productivity and decrease of bioprocess development times are requested from an economic point of view. In addition, the regulatory authorities strongly suggest science-based development of biopharmaceutical processes. One

of the most prominent regulatory initiative of the last decade was the process analytical technology (PAT) and quality by design (QbD) initiative of the American Food and Drug Administration (FDA) [3,4]. Main goals of PAT and QbD are to understand and control manufacturing by focusing on the design of the process and thereby to consistently ensure a predefined quality at the end of the process [5].

In the last decades, multivariate data analysis (MVDA), mechanistic modelling and dynamic experiments emerged as efficient tools to achieve faster process development, higher productivity and a high degree of process understanding as requested by the regulatory authorities. Using MVDA, a broad spectrum of statistical tools are applied to identify major influential process parameters on quality attributes such as multivariate linear regression and dimensionality reduction tools, which are reviewed in detail elsewhere [6,7]. Mechanistic process modelling is successfully applied for fast and efficient process optimization, for example, the optimization of upstream fermentation processes [8]. Typically, the model structure is initially formulated applying first order principles, such as a set of differential equation of mass or energy balances, one for each chemical species. Unknown parameters, e.g. kinetic reaction rates, of these so called state equations can be either directly estimated from measured states, e.g. by rearranging Eq. 1, or found by varying the parameters in a optimization procedure aiming to minimize the difference between the state estimates and the measured states. Dynamic methods refer to the purposeful dynamic deflection of process states with the goal of maximizing the information extractable from an experiment. For example, a dynamic experiment may deal with the cellular response to rapid changes in the feed profile or environmental conditions (pH, T) carried out either in a continuous stirred tank reactor or under fed-batch mode. This approach was successfully used for the identification of optimal feeding profiles [9] and the characterization of induced state metabolic capabilities [10].

It has been shown that the usage of specific metabolic turnover rates as input for MVDA is beneficial compared to classical approaches using raw process data, since the calculation of rates introduces additional process knowledge as a feature extraction step [11,12]. In mechanistic modelling, specific rates play a major role as parameters, which establish relationships between state equations. Even on a more detailed level, specific rates are inputs for metabolic relationships for flux analysis [13]. Moreover, the estimation of specific rates is key to the evaluation of dynamic experiments, where the impact of a collection of parameter combinations on the process quality can be investigated within one process run [14,15].

As outlined above, a key element of a broad spectrum of tools for efficient bioprocess characterization, optimization and investigation is the calculation of volumetric ($r_i$) and thereof derived specific turnover rates ($q_i$). In principle, this is performed based on material balances, as depicted in Eq. 1 and Eq. 2 [16]. The basic idea is that the change of mass of a species $i$ within a bioreactor system with a given reactor volume ($V_R$) can only occur due to inflows ($\dot{V}_{in} \cdot c_{in}$) and outflows ($\dot{V}_{out} \cdot c_{out}$) with their respective concentration as well as the reaction ($V_R \cdot r_i$) within the system. In order to obtain catalyst specific rates, the volumetric rates ($r_i$) are divided by the biomass concentration ($\gamma_X$).

$$\dot{V}_{in} \cdot c_{in} - \dot{V}_{out} \cdot c_{out} + V_R \cdot r_i = \frac{d(V_R(t) \cdot c_i(t))}{dt} \quad (1)$$

$$q_i = \frac{r_i}{\gamma_X} \quad (2)$$

In an ideal steady-state operation mode of bioreactors the volumetric and specific turnover rates are constant over time. This process mode opens up advantages compared to batch and

fed-batch processes such as large scale production of cheap products at relatively high productivity. However, a continuous design of fermentation processes holds also risks such as strain instability due to mutations on the long run, which limits its application in the production of biologics in industry [17,18]. Therefore, batch and fed-batch processes are still a dominant production form, where the dynamic changes of turnover rates over time are essential for the description and understanding of the process. Herein, this contribution will focus on batch and fed-batch processes.

The concise biotechnological formulation of material balances has been investigated over the last three decades. In the case of no in and outflow of the system material balances are reduced to a simple derivate formulation ($r_i = \frac{d(c_i(t))}{dt}$). If in- and outflows are present, the material balance can be integrated and the resulting cumulative entity is differentiated as shown by Herwig [14]. Especially the calculation of turnover rates of gaseous species is more sophisticated since the dilution of the off-gas stream due to water stripping has to be taken into account [19]. From a formal point of view, a biotechnological system is characterized by the number of turnover rates, which can be estimated directly from measurements and those which may be calculated from others [20]. This classification can be used to identify redundant equations, which subsequently can be used for reconciliation and gross error detection. This approach significantly increased the signal to noise ratio on the obtained reaction rates in industrial fed-batch fermentations [21].

A major aspect of rate calculation is the pre-processing of the original process data ($V(t), c_i(t), \dot{V}_{in}, \dot{V}_{out}$) in order to derive smooth and interpretable rates. Currently, this is performed by data driven smoothing and filtering algorithms, such as averaging and polynomial smoothing, low pass frequency filters as well as model based filters. Digital low pass filters (e.g. the commonly applied Butterworth filter) and smoothing algorithms (e.g. Savitzky-Golay smoothing) are applicable for the filtering of online signals with small sampling intervals and thereof derived rates, since their smoothing ability relies on the vast amount of sampled data [22–24]. On the other hand, these approaches are not suitable for the filtering of rates based on offline signals such as metabolite formation rates, growth rates and product formation rates. For the filtering of rates derived from offline signals, filters based on mechanistic process models such as Kalman filters or particle filters can be used [25].

Recently, the precision of rates as a function of the precision of the originating signal was investigated and it was found that high sampling frequencies of the integral signal lead to low signal to noise ratios (SNR) on the calculated rates [16]. This is especially important, since a uniform SNR on the derived rate is prerequisite for a processing of rates within MVDA tools, such as multi-linear regression, principal component regression and partial least squares regression models. Non-uniform SNR ratios and herein not homoscedastically distributed regression residues can lead to imprecise multivariate models and thereby to possibly wrong predictions [26].

Hence, for the accurate volumetric and specific rate calculation following goals can be summarized:

i.) A concise formulation of the material balances around the system boundaries must include all relevant in- and outflows as well as the classification of calculable and balanceable conversion rates for the whole process system [16,19,20].

ii.) Optimal pre-processing of data is necessary in order to obtain interpretable rates at reduced noise level.

iii.) A uniform signal to noise ratio (SNR) on all resulting rates is required for a valid comparison and for most multivariate tools [16].

However, to date, no method has been published that enables to derive rates with uniform SNR at varying sampling frequencies and sampling precision of the integral signal ($V(t) \cdot c_i(t)$). Furthermore, current empirical rate calculation strategies are prevailing and no generic method to identify a workflow consisting of pre-processing and rate calculation steps in order to derive most precise and robust rates has been published.

In this contribution, we overcome the aforementioned gaps by presenting an advanced novel algorithm that is capable of delivering precise and accurate volumetric rates with uniform SNR at dynamic process behaviour and varying sampling frequencies. Furthermore, we present a generic model-based algorithm for deriving an optimal rate calculation workflow of pre-processing, rate calculation and post-processing steps. The power of the presented tools is demonstrated on industrial microbial processes validated with *in silico* and real process data.

# 2. Materials and Methods

## 2.1 Generation of model based in silico data

An empirical model adapted from Sonnleitner [27] was used to describe a representative microbial fed-batch fermentation with oxidative growth and one carbon feed as described in detail in section S1 of the supporting information (section 9.1). Although this is a very simple model of a fed-batch cultivation, the introduced dynamics in the turnover rates are suitable to compare different rate calculation algorithms in respect to their capability of de-noising and retaining the dynamics. Gaussian distributed noise relative to the amplitude to the power of 1.05 was added to the modelled biomass signal, mimicking an increased relative error at higher cell densities. This is frequently observed, since higher viscosity of the media requires more dilution steps - each connected to a dilution error - in case of biomass estimation via OD measurement. The modelled feed scale signal was superimposed with an absolute error of 0.05 to 0.5 g, which is regarded as a typical weight scale error level. In order to obtain statistically significant results the addition of noise and subsequent rate calculations were performed 50-fold. The sampling of the in silico generated biomass was non-equidistant and adapted from real process sampling intervals [28]. The sampling interval of the feed scale was set as 20 s.

## 2.2 Real process data for verification of the optimal rate calculation workflow

For the verification of the established workflows, two representative fed-batch fermentations using the model organism *Pichia pastoris* producing recombinant horse radish peroxidase were analysed. Fermentation conditions are described in [29]. The weight scale signal as well as the off-gas measurements of $CO_2$ were considered as online signals whereas the dry cell weight measurement (DCW) was considered as a typical offline measurement for the analysis with the proposed rate calculation algorithms. No additionally recorded real process data was necessary to investigate the closing of the carbon balance after the application of different rate calculation workflows.

## Computational methods

Simulations and rate calculation workflows were implemented in MATLAB® 2015 (The Mathworks, Natick, MA, USA) using the InCyght bioprocess technology toolbox v1.0 (Exputec, Vienna, Austria).

## 2.3 Material balances

The formulation of material balances and the corresponding calculation of the volumetric biomass formation rate ($r_X$), substrate uptake rate ($r_S$) and carbon emission rate (CER) were performed as described in detail elsewhere [16].

## 2.4 Pre-processing methods

### Butterworth low pass filter

Butterworth filters are low pass filters which reject unwanted high frequencies at a maximally flat response in the passband first published by Butterworth [30]. A significant advantage of the Butterworth filter compared to other low pass filters is that it has a flat frequency response in the passband and a smooth roll-off towards zero in the stopband. A practical limitation with non-periodic signals, as usually present in biotechnological applications, is that fluctuations might be introduced at start and end points of a time series. Computational implementation of the Butterworth filter design were performed by using the MATLAB® function *butter*. For estimating the filter coefficients the algorithms described by Parks were used [31].

### Savitzky-Golay smoothing and first derivative

The Savitzky-Golay filter is one of the most used and ubiquitous polynomial filters. It is based on the approximation of the data by a polynomial with specified degree fitted to a subset of the data points within a certain window. Once the polynomial function is known it is easy to calculate derivatives of the smoothed signal. A common drawback of this filter is that the window size is fixed over the entire signal insensitive to dynamic changes. Savitzky-Golay smoothing and calculation of the first derivate was done by using the MATLAB® function *sgolay*. Implementation was carried out as described by Orfanidis [32].

### Extended Kalman filter algorithm

Kalman filters, first reported by Kalman [33], are minimal-variance estimators that produce statistically optimal estimates of the states of a system using a dynamic model. Extended Kalman filters are the non-linear extension of classical Kalman filters and well known applications are navigation systems, signal filtering but also bioprocess control. Here, an extended Kalman filter configuration was used which is described in detail in section S2 of the supporting information (section 9.1).

### Statistical test values for comparing time signals (NRMSE)

As a statistical test value for the comparison of the calculated rate by different algorithms to the true rate the normalized root mean square error (NRMSE) was applied according to:

$$NRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{n}(r_{calc,i}-r_{true,i})^2}{n}}}{r_{true,mean}} \quad (3)$$

This measure was used since strong deviations from the true rate are penalized more seriously compared to other measures, such as the percentage of error or SNR.

# 3. Results

## 3.1 Method Development

### 3.1.1 Algorithm to calculate metabolic turnover rates with constant signal to noise ratio: Dynamic window rate calculation algorithm

Here, we describe a novel algorithm to calculate metabolic turnover rates with constant signal to noise ratio.

25

The uncertainty boundary of the rate ($U_r$) and the SNR of the rate can be calculated with rules of Gaussian error propagation resulting in Eq. 4A, Eq. 4B and Eq. 5. $U_{y_i}$ corresponds to the uncertainty boundary, e.g. obtained due to an investigation of historical data, of a data point $i$ of the originating signal and $\Delta t$ corresponds to the span of the time window. If the measurement uncertainties are estimated by standard deviations, e.g. from replicate measurements, Eq. 4B can be used instead to derive the simple measurement uncertainty of the rate $u_r$. In the following sections we assume knowledge about the uncertainty boundaries of the input signals and therefore apply 4A. However, it should be noted that all presented algorithms can also be formulated alternatively with Eq. 4B. The full derivation leading to those equations is shown in section S3 of the supporting information (section 9.1).

$$U_r = \frac{U_{y_1}+U_{y_2}}{\Delta t} \text{ (4A)}$$

$$u_r = \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\Delta t} \text{ (4B)}$$

$$SNR_r = \frac{r}{U_r} \text{ (5)}$$

Based on these equations, an algorithm can be formulated to estimate turnover rates with constant signal to noise ratios, as depicted in Figure 1.

- Step 1: The Algorithms starts at the first data point (j = 1) and calculates the rate to the second point, which equals a window size of one point (i = 1). In case of continuing this procedure over the whole signal, we would call the algorithm fixed window rate calculation using two points (FWR2).
- Step 2: The SNR of the resulting rate with the current window size i is calculated using Gaussian error propagation (Eq. 5).
- Step 3: If the calculated SNR of the rate does not meet a desired SNR level set by the operator, the window is widened by one point (i = i + 1) and the algorithm continues from step 2. As soon as the calculated SNR of the rate fulfils or exceeds the requirements, the calculated rate together with its SNR is stored and the algorithm proceeds with step 1 setting j = j + i.

This procedure is continued until the end of the signal. A requirement for the algorithm is the availability of data points of a process signal and their corresponding standard deviation or SNR.
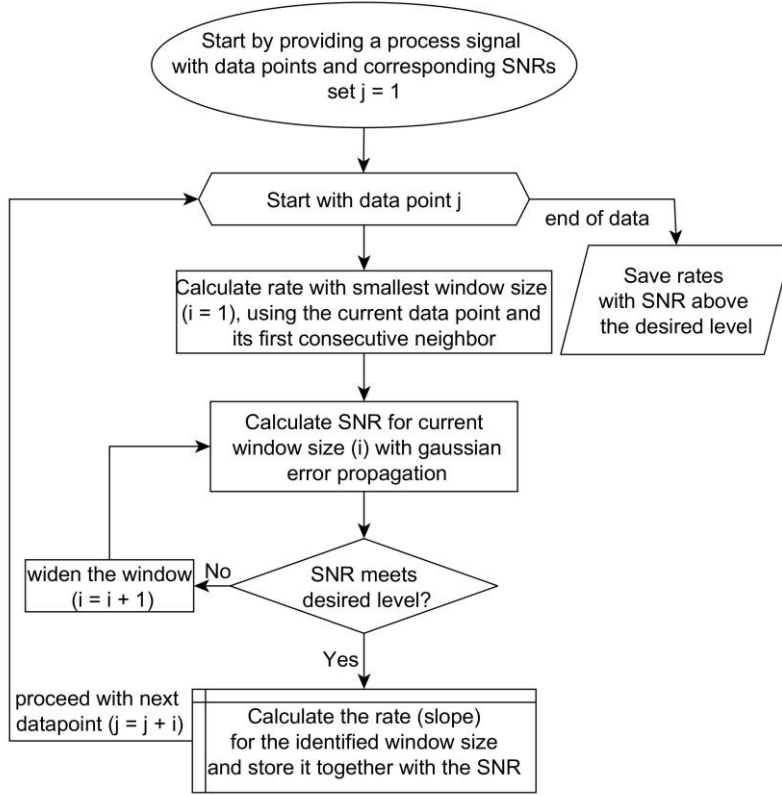
*Figure 1: Schematic representation of the dynamic window rate calculation (DWR) algorithm. Step 1: The rate between the first two points is calculated, which equals a window size of i = 1. Step 2: The SNR of the rate is calculated using Gaussian error propagation using Eq. 4A and Eq. 5. If the SNR is higher than desired, the rate is stored and the algorithm proceeds with the next rate starting at data point j = j + i. If the desired SNR is not met the window is widened by one data point (i = i + 1) and the algorithm repeats from step 2.*

The performance of the developed algorithm on in silico data and real-process data is described in section 3.2.1 and 3.2.2 and visually displayed in Figure 3C.

### 3.1.2 Generic algorithm for identifying an optimal rate calculation workflow

Next to an algorithm delivering rates with a uniform signal to noise ratio (section 3.1.1), it is necessary to develop a generic algorithm capable of identifying optimal rate calculation workflows including data pre-processing, rate calculation and post-processing steps. As a final goal this workflow should deliver turnover rates, which have high accuracy and high precision insensitive to noise in the originating signal. To do so, a novel model-based approach is proposed: As a first step, a mechanistic or empirical model is developed mimicking most precisely the states of a given process, which are subsequently superimposed by a representative noise. We want to note that the addition of noise influences which rate calculation workflow may be considered as optimal. Ideally, the real-process signal noise should be added to in silico generated data. However, this is generally not known. Therefore, a good noise approximation is required and deep understanding of the true nature of the signal noise is crucial.

After the application of different combinations of pre-processing, rate calculation and post-processing steps, the NRMSE between the calculated rate and the true rate shows the power and applicability of a chosen workflow (Figure 2). However, due to the differences in sampling frequency, online signals (continuous sampling < 5min) and offline signals (few samples over the whole process) have to be considered separately. Most crucial is the careful selection of

27

filter parameters, such as cut-off frequencies of frequency filters or order of polynomials and window width of smoothing algorithms. Therefore, filter parameter selection according to rational aspects are supposed to lower the risk of deteriorating the signal characteristics and are presented in section S4 of the supporting information (section 9.1).

Reducing noise of the originating signal during the rate calculation process itself can be achieved by widening the temporal window applied for estimating the slope. Moreover, algorithms can be differentiated whether they use a fixed (FWR), a dynamically adapted window (DWR) size or a window spanning between the boundaries of processes phases (PBR).
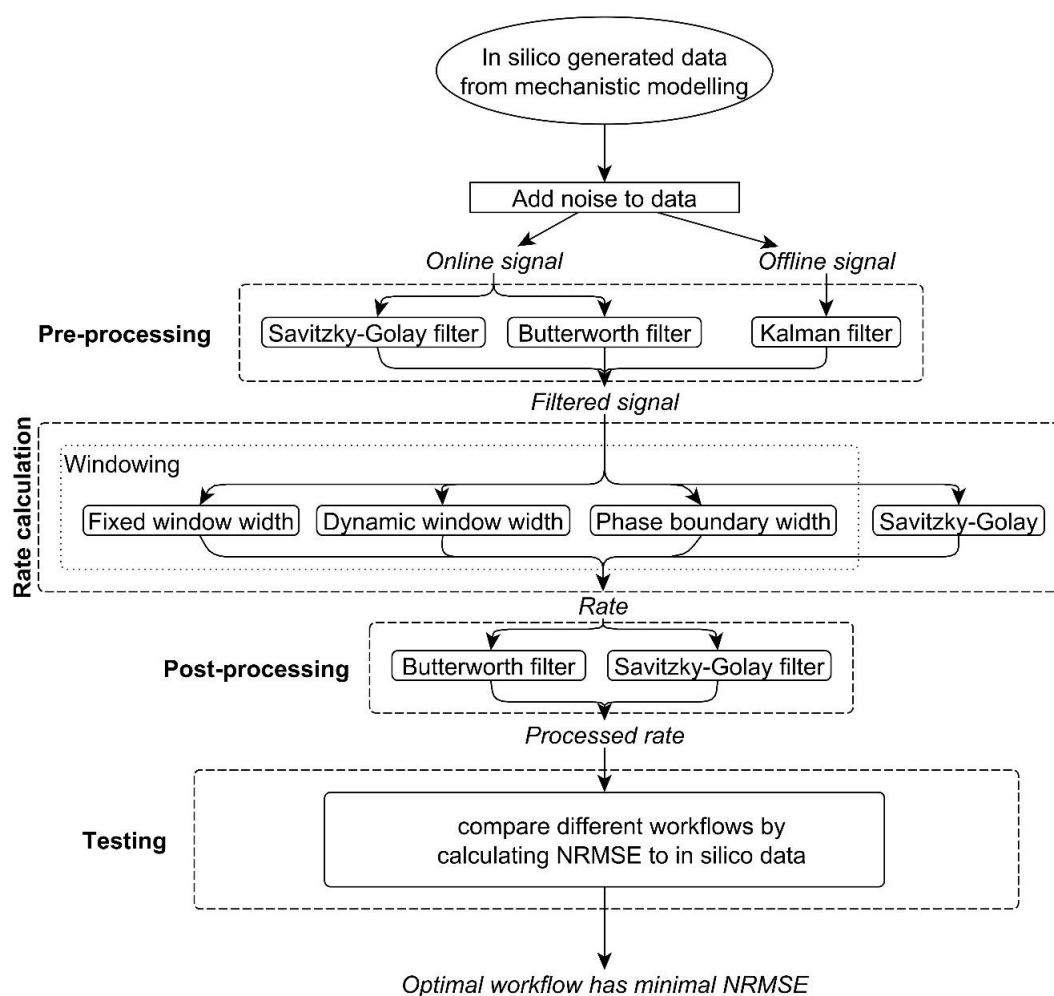


Figure 2: Graphical representation of the generic method to identify an optimal rate calculation workflow consisting of pre-processing, rate calculation and post-processing steps. First of all, model generated data is pre-processed using rationally derived filter parameters as described in section S4 of the supporting information (section 9.1). For online signals, frequency filters (e.g. Butterworth filter) and polynomial smoothing algorithms (e.g. Savitzky-Golay filter) are suggested whereas for offline signals model based Kalman filters are used. Subsequent rate calculation can either be performed using fixed windows, dynamic windows or windows spanning between phase boundaries. Fixed window rate calculation with two data points is the most frequently used way of calculating the derivate. Another way of estimating the slope is to differentiate a polynomial which fits the data. This is done by taking the first derivative of the Savitzky-Golay polynomial. Finally the derived rates can be post-processed with frequency or polynomial filters, which is in this contribution again only applied for rates derived from online signals using rationales explained in section S4 of the supporting information (section 9.1).

28

## 3.2 Application of Presented Algorithms

### 3.2.1 Developing optimal rate calculation workflow for microbial fermentation processes

In silico generated fed-batch data is used to identify the optimal rate calculation workflow for microbial fermentation processes applying the generic algorithm presented in section 3.1.2.

We consider one representative rate derived from an online and offline signal, respectively. As a classical offline signal the biomass measurement was chosen, which leads to $r_x$ including the assumptions of no inflow and outflow of biomass. As a typical online measurement the weight scale signal of the carbon feed was chosen, which led to $r_S$ assuming carbon limited fermentation [16]. The performance of selected workflows in respect to the precision and accuracy of derived $r_x$ and $r_S$ is shown in Figure 3 and Figure 4, respectively.

As shown in Figure 3A, the combination of Kalman filtering (KaF) and dynamic window adaptation with a desired SNR of 5 (DWR5) for $r_x$ was identified as optimal workflow followed by the combination of KaF and fixed window rate calculation with a window size of two points (FWR2). FWR2 is the common way of calculating a rate between two points [14,16,34]. The superior behaviour of KaF + DWR5 in comparison to KaF + FWR2 can be explained by the fact that Kalman filters not only improve the estimation of the true signal but also provide an improved estimation for the covariance of the signal according to Eq. 9 of the supporting information (section 9.1). As shown in Figure 5, this additional information can subsequently be used by DWR5 to dynamically widen the window size for rate calculation and thereby gain even more significant rates with uniform SNR. Surprisingly, DWR5 alone was nearly as good as the combination with Kalman filtering as depicted in Figure 3A, which remarkably underlines the power of this algorithm taking into account that no mechanistic model is required. The rate calculation using solely FWR2 is always worse than all other rate calculation algorithms for noise levels higher than 1% relative error. The precision (relative error) of $r_x$ calculated with different algorithms is shown in Figure 3B, which equals 1/SNR. Rate calculation using windows equal to the phase boundaries (PBR) had the lowest relative error for noise levels lower than 5%. Upon this boundary the combination of Kalman filtering and dynamic window adaptation rate calculation (KaF + DWR5) showed higher precision. The relative error of FWR2 ranged from 2 to 7, depending on the noise level, which indicates the extremely low reliability of this rate calculation method.
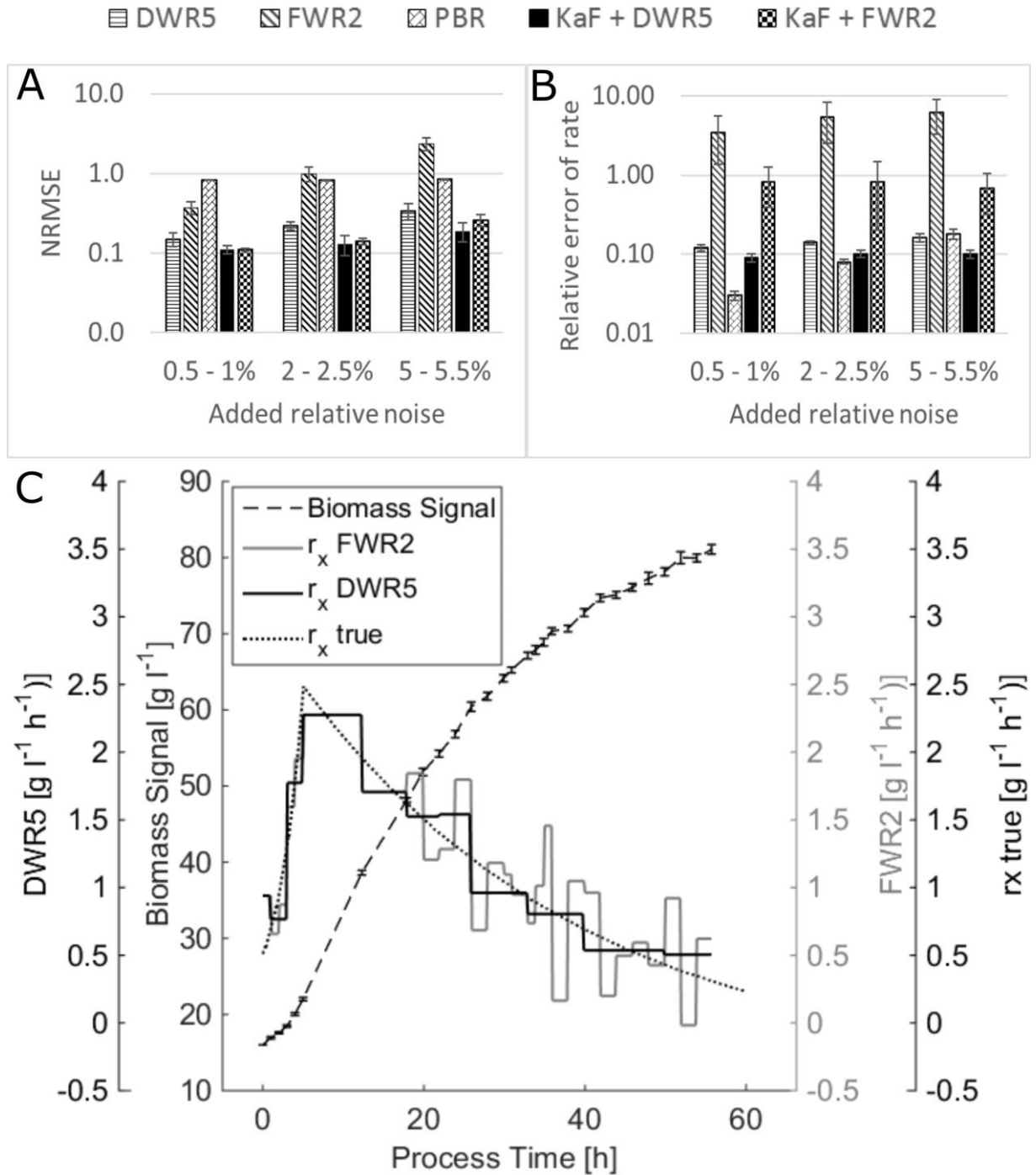
*Figure 3: A: comparison of different rate calculation algorithms for the volumetric biomass formation rate ($r_x$) originating from an simulated OD-measurement (offline signal) by means of NRMSE to the true rate at different relative noise levels. Error bars indicate 1 standard deviation from 50 simulations. B: relative errors on rate ($r_x$) obtained by different rate calculation workflows with different relative noise levels. Error bars indicate 1 standard deviation from 50 simulations. C: time resolved biomass signal with 2% relative error indicated with error bars and derived rates ($r_x$) using algorithms FWR2 and DWR5 in comparison to the true $r_x$. The stairs function for the rates indicate validity ranges. Used abbreviations for algorithms: DWR5: rate calculation with dynamic window adaptation with SNR of 5 on resulting rates, FWR2: rate calculation with fixed window size of 2 points, PBR: window size equals phase boundary, KaF: Kalman filtering.*

For online signals, the combination of the first derivate calculation by Savitzky-Golay algorithm (SGR20) and a subsequent Butterworth low-pass filter with a cut-off frequency of 0.5 mHz (BwF0.5) resulted in an similar accurate rate as the estimation with Savitzky-Golay smoothing (SGF20), dynamic window adaptation rate calculation (DWR5) and a final low-pass Butterworth filter (BwF0.5) as shown in Figure 4A. It was not possible to detect a significant difference (two sample t-test with significance level $\alpha = 0.05$) between those two approaches for all noise levels. A possible interpretation is that the applied algorithms filter noise in a redundant way and therefore addition of DWR5 does not significantly decrease the noise level. Nevertheless, the approach with SGR20 and BwF0.5 seems to be beneficial since it consists of only two algorithms and therefore minimizes the risk of signal deterioration caused by an additional de-noising algorithm. However, the current implementations of the Savitzky-Golay first derivative do not provide an estimate for the precision of the resulting rates, therefore the relative error is not shown in Figure 4B for the combination SGR20 + BwF0.5. The relative error of algorithm combinations containing DWR are always lower (< 0.2) than rate calculations with fixed window size (> 7) as compared in Figure 4B.
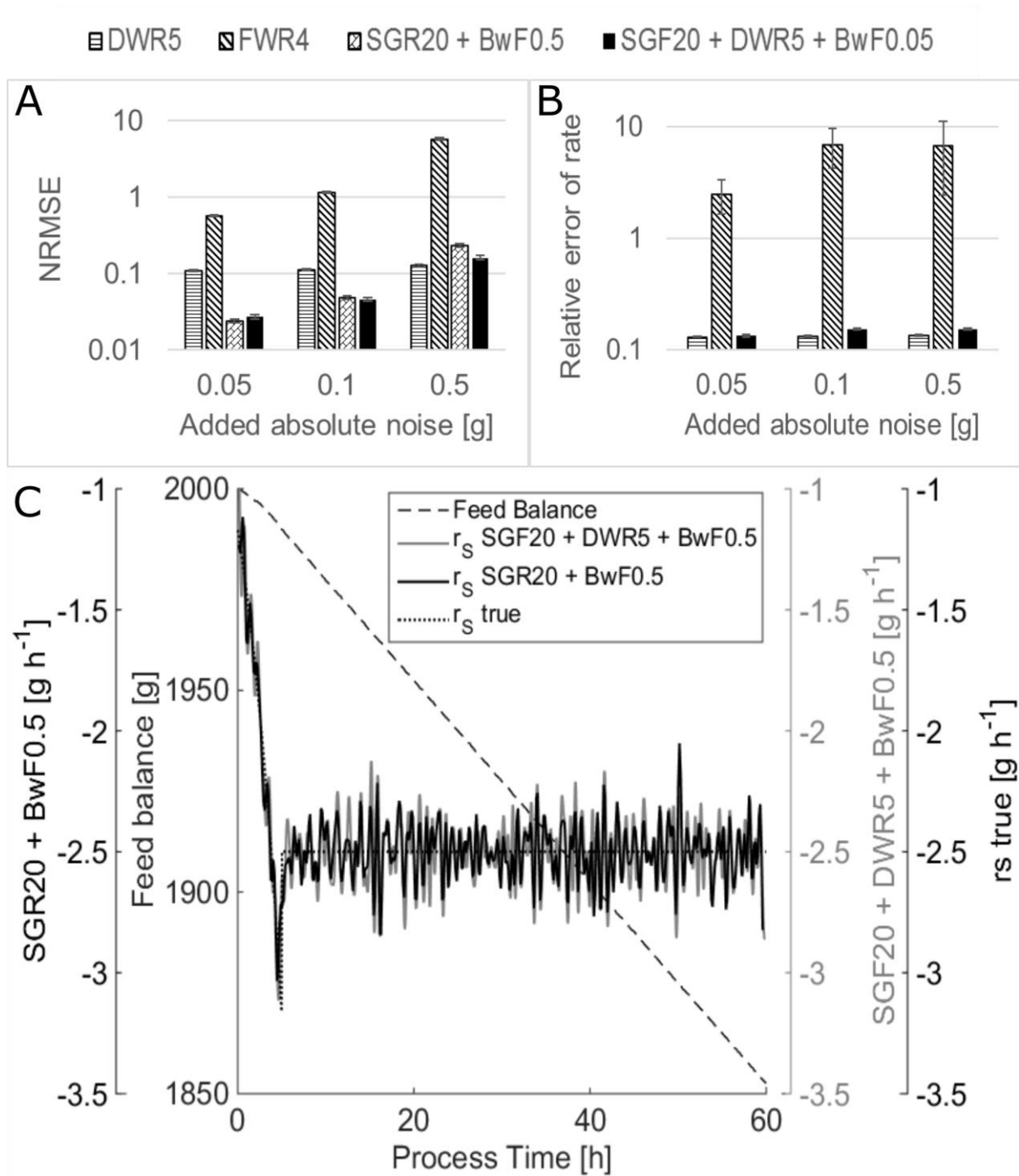
*Figure 4 A: comparison of different rate calculation algorithms for the substrate uptake rate ($r_S$) originating from a simulated scale signal (online signal) by means of NRMSE to the true rate at different absolute noise levels. Error bars indicate 1 standard deviation from 50 simulations. B: relative errors on rate ($r_S$) obtained by different rate calculation workflows at different absolute noise levels. Error bars indicate 1 standard deviation from 50 simulations. C: time resolved feed scale signal with 0.1 g absolute error and derived rates ($r_S$) using algorithm SGR+BW0.5 and DWR5 + SG in comparison to the true $r_S$. Used abbreviations for algorithms: DWR5: rate calculation with dynamic window adaptation with SNR 5 on resulting rates, FWR4: rate calculation with fixed window size of 4 points, SGF20: Savitzky-Golay smoothing with a window size of 20 data points, SGR20: Savitzky-Golay first derivative with a window size of 20 data points, BwF0.5: Butterworth low pass filter with 0.5 mHz cut-off frequency.*
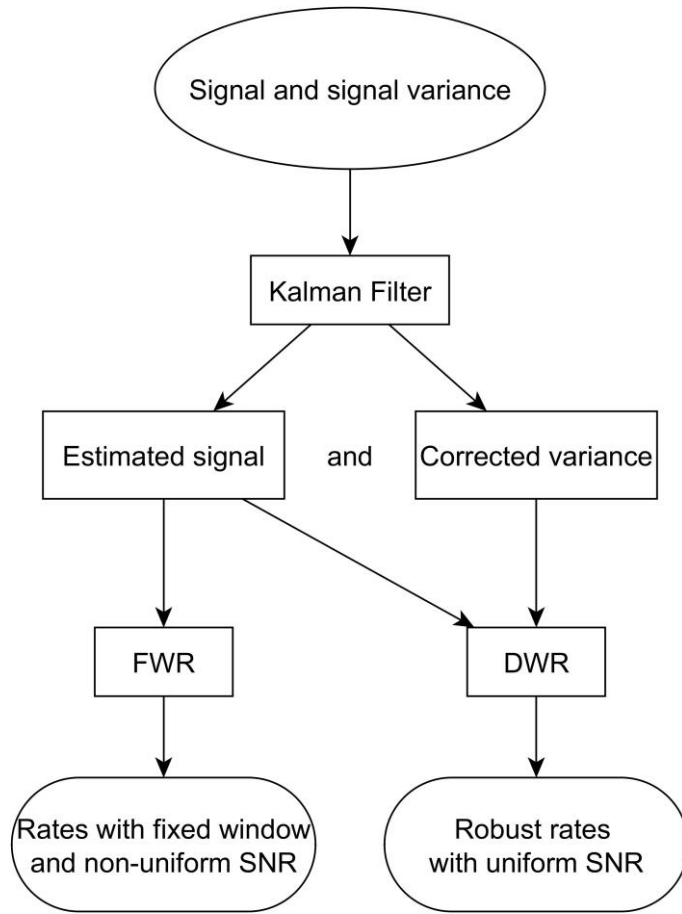
*Figure 5: Schematic workflow and synergy between Kalman filter and dynamic window rate calculation (DWR). Kalman filter provide an estimation of the variance, which can be used by DWR to dynamically widen the windows for rate calculation. Thereby the precision of the resulting rates can be increased and uniform SNR can be achieved. Rate calculation with fixed windows (FWR) does not use the corrected variance as input and therefore less significant rates with non-uniform SNR evolve.*

More combinations of algorithms for deriving $r_x$ and $r_S$ are shown in Table 1 and Table 2 of supporting information (section 9.1). The optimal rate calculation workflows for online and offline signals for a standard microbial fed-batch fermentation are summarized in Figure 6.
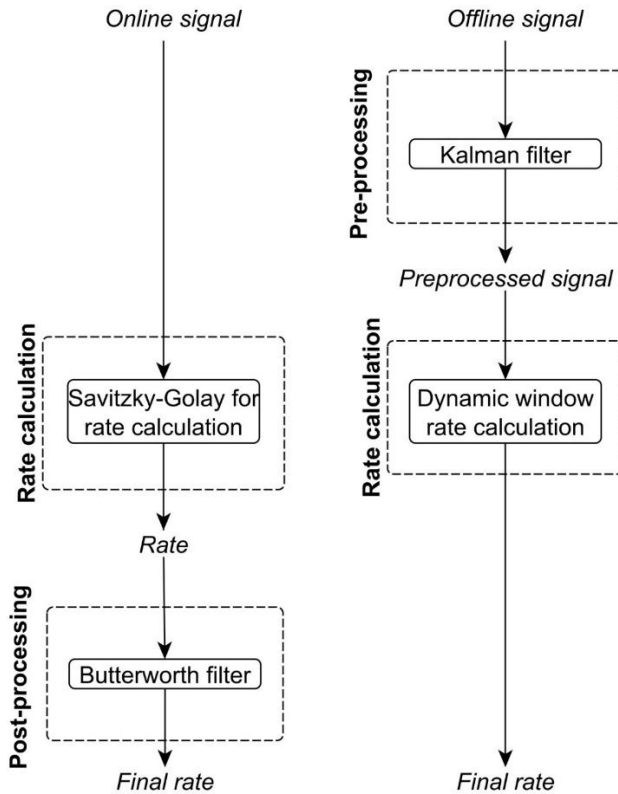
*Figure 6: Proposed optimal rate calculation workflow for online and offline signals consisting of pre-processing, rate calculation and post-processing steps for obtaining most precise and robust rates in standard, industrial microbial fermentations.*

### 5.2.2  3.2.2 Proofing optimal microbial rate calculation workflow on real process data

Since the resulting workflows for a standard microbial fed-batch fermentation established by *in silico* generated process data in section 3.2.1 are biased by the nature of the added noise, we need to verify the optimal workflows for online and offline signals using real process data. As a typical microbial fed-batch fermentation, we used data from Pichia pastoris fed-batch fermentation with methanol induction. As typical for many biopharmaceutical processes, a mechanistic model is not available and therefore workflow options with Kalman filtering must be neglected.

Again, the generic algorithm presented in section 3.1.2 can be used for optimal workflow identification. For real process data, the closing of the C-balance can be used as a measure for precision (NRMSE) since true values of rates are not available.

In analogy to the *in silico* observations, the calculation of $r_x$ using dynamic window adaptation (DWR5) showed best performance concerning the dynamic of the signal, excluding outliers with low SNR in the originating signal (Figure 7C). This led to a major improve in the closing of the C-balance (increase of 50%-400%) compared to all other workflows without DWR5 calculation of $r_x$ (Figure 7A). The usage of the optimal workflow for $r_S$ suggested by the in silico observations for normal noise on balances < 0.1 g (SGR20 + BwF0.5) had major improve concerning the smoothing of $r_S$ (Figure 7B), but resulted only to minor improves concerning the closing of the C-balance (Figure 7A).

In general, the *in silico* identified optimal workflows for $r_x$ and $r_S$ were found to show the best closing of C-balances for real process data. This proves the applicability of the optimal rate calculation workflow for microbial fermentations.
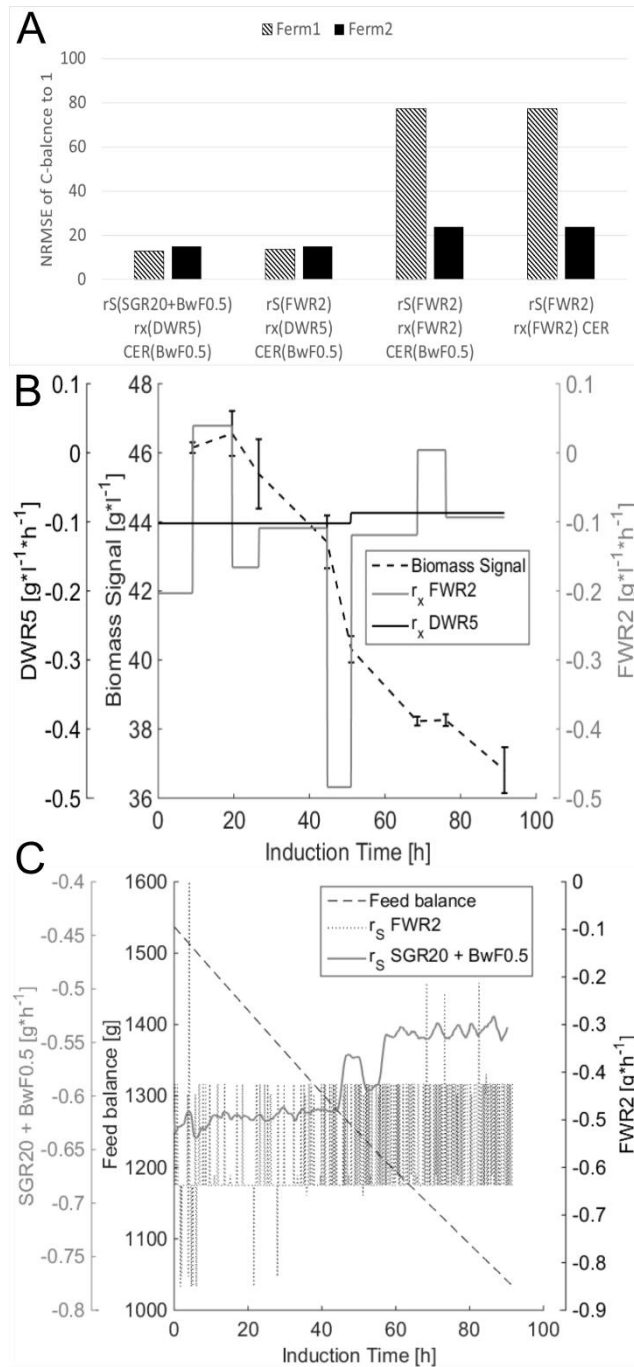


*Figure 7: A: Comparing different rate calculation workflows for $r_X$, $r_S$ and CER originating from real process signals by calculating the NRMSE of the resulting C-balances to 1. B: Time resolved feed scale signal and derived $r_S$ using algorithms FWR2 and SGR20 + BwF0.5. C: time resolved biomass measurement and derived $r_X$ using algorithms DWR5 and FWR2. The stairs function for the rates indicate validity ranges. Used abbreviations for algorithms: DWR5: rate calculation with dynamic window adaptation with SNR of 5 on resulting rates, FWR2: rate calculation with fixed window size of 2 points, SGR20: Savitzky-Golay first derivative with a window size of 20 data points, BwF0.5: Butterworth low pass filter with 0.5 mHz cut-off frequency.*

# 4. Discussion

## 4.1 Comparison of DWR algorithm to different rate calculation methods

In signal analysis the result of a measurement or a reconstruction of a signal is an estimate of the true signal and therefore accompanied by an error. This error consists of two components, a systematic error, determined by the accuracy, and a stochastic error or uncertainty, described by the precision. The normalized root mean square error (NRMSE) is the most frequently used marker throughout many disciplines for signal accuracy since it reflects the deviation of the mean of the de-noised signal to the true signal [35,36]. The signal to noise ratio (SNR) reflects the lack of exact knowledge of the value and is therefore a frequently used estimate of the precision [37]. Therefore, the reconstructed signal with the lowest NRMSE and SNR is regarded as the most precise and robust estimation.

In literature, a broad spectrum of approaches to calculate rates in dependence on the bioprocess investigated are reported, for example, based on finite difference approximation or direct calculation of specific rates with the integral of viable cells [38]. However, we want to note that all approaches rely on the mathematical calculation of the slope. This can be reduced to a numerical finite difference approximation problem (Eq. 13 in the supporting information in section 9.1), unless a priori knowledge about the signal characteristics is available (e.g. exponential growth). Therefore, we anticipate that the dynamic window adaptation algorithm (DWR) is fully applicable to all kind of rate calculations with linear signal characteristics and superior to other rate calculation methods as shown in Figure 3 and Figure 4, independent of the underlying organism or signals processed. The essential difference between DWR and other algorithms relies on the sophisticated way how data points of the integral signal are selected to derive sound rates but not on the way how the derivative itself is formed.

## 4.2 Applications of DWR algorithm and the generic algorithm for deriving an optimal rate calculation workflow

As outlined in the introduction within a system of calculable and measurable rates redundancy can be used in order to detect gross errors and improve data quality by reconciliation [20,39]. Testing the significance of errors of the measured rates is a crucial step within the reconciliation procedure. This is done by calculating a statistical test value ($h_{\varepsilon}$), which requires knowledge about the precision of the measured rates. To the best of our knowledge, the estimation of errors on measured rates was so far conducted only by empirical approximations. The presented DWR algorithm can be used to derive rates and their associated precision and subsequently formulate the covariance matrix of measured rates. Thereby, we anticipate that the overall accuracy in terms of closing of material balances can be increased and the precision of calculated and reconciled rates can be improved, although this requires further investigations of the interactions between DWR and reconciliation.

Another potential application of DWR is the evaluation of dynamic experiments such as dynamic design of experiments, where the goal is to explore multiple system responses within a single process run. Although the temporal resolution of calculated rates cannot be refined with DWR, since only widening of time windows is possible, the algorithm ensures precision on the observations, reducing misleading interpretations due to temporal system fluctuations.

Besides the offline applications of DWR for process data analysis, online operation can be used to perform a stable process control triggered by reliable and precise specific rates.

## 4.3 Applications of DWR algorithm and the generic algorithm for deriving an optimal rate calculation workflow in combination with MVDA

Many MVDA correlation methods are built on the assumption that the resulting regression must attain homoscedastic distribution of residues [26]. Due to changing signal dynamics and a varying precision of the measurement, rates with non-uniform SNR are obtained. If those rates are subsequently correlated to a critical quality attribute (CQA), homoscedastic residues cannot be achieved and outliers may alter the correlation significantly. Therefore, we anticipate that rates with uniform SNR as provided by DWR lead to more significant correlations which are robust against outliers in the original process signal, as shown in Figure 7C.

Mechanistic models use specific rates and derived entities (e.g. yield coefficients) as physiological parameters, which connect process states with each other. For their estimation, DWR is anticipated to deliver more precise and robust inputs which reflect physiological interactions more accurately.

## 4.4 Extensions of DWR algorithm and the generic algorithm for deriving an optimal rate calculation workflow to mammalian cell cultures

The DWR algorithm and the optimal rate calculation workflow in their presented form are tested with batch and fed-batch data. The optimal rate calculation workflow might slightly change for different biotechnological application depending on the dynamics of the turnover rates. For example, higher dynamics might require smaller windows for the Savitzky-Golay smoothing. However, the proposed generic approach which delivers optimal rate calculation workflows as presented in 3.1.2 is applicable to all biotechnological applications. Moreover, the presented dynamic window rate calculation algorithm is anticipated to greatly improve rate calculation in mammalian batch and fed-batch processes, where noise prohibits the observability of true rate changes due to the slow growth characteristics of the process. Here, a different specific rate calculation method for mammalian cell culture than for microbial fermentations is commonly applied. In this slightly different rate calculation procedure it is suggested to use the ratio of the observed mass difference during a time window $(t - t_0)$ of species $i$ $(\Delta m_i)$ to the integral of formed viable cells $(IVC)$ during that time window according to Eq. 6 and 7 [38,40]. As shown in section S6 of the supporting information (section 9.1), the estimation of $q_{i,I}$ using this different approach (Eq. 7) is equivalent to the finite difference approximation and subsequently dividing the rate by a linear approximated cell concentration, which was used throughout this contribution to drive specific rates. Hence, the algorithms proposed here seem to be highly promising also for mammalian cell culture.

$$IVC = \int_{t_0}^{t} \gamma_{X_V} \cdot dt \ (6)$$

$$q_{i,I} = \frac{\Delta m_i}{IVC} \ (7)$$

## 5. Conclusions

The calculation of robust and precise turnover rates for univariate and multivariate bioprocess analysis, requires i) a concise formulation of material balances in respect to the system boundaries, ii) optimal data pre-processing of process data and post-processing of rates and iii) a uniform signal to noise ratio (SNR) on all resulting rates.

Optimal data pre-processing of process data and post-processing of rates can be greatly improved by the presented dynamic window adaptation algorithm (DWR). This algorithm is

37

based on the dynamic widening of time windows of the derivative, triggered by the signal-to-noise ratio (SNR) of the integral process signal. The algorithm ensures uniform SNR on the resulting rates as required by MVDA tools (e.g. multi-linear regression) and thereby delivers more precise and robust rates against outliers as demonstrated by in silico and real process data. The algorithm is generically applicable for both the analysis of historical data sets and real-time applications, and was tested for batch and fed-batch processes in this contribution

Optimal rate calculation workflows can be identified using the presented generic model-based algorithm, which requires a process model and assumption about the signal errors.

For industrial microbial batch and fed-batch processes, an optimal rate calculation workflow was identified for offline and online signals. For online signals, a combination of Savitzky-Golay first derivative followed by a Butterworth filter on the rate led to optimal rates concerning precision. For optimal rate calculation starting with offline signals, Kalman filters are used to pass the estimated variance of the signal to the DWR, which subsequently optimally adapts the window size.

The presented methodology to calculate precise rates is extendable to all microbial bioprocesses and can be assumed to be beneficial for mammalian cell cultures.

We anticipate that using the presented optimal rate calculation workflow, including DWR, will improve the performance of MVDA tools and mechanistic modelling and thereby increase the overall competitiveness and overall bioprocess understanding.

## Acknowledgements

## Symbols Used

Symbols

| | | |
|---|---|---|
| $c_{in}$ | [g l-1] or [mol l-1] | Concentration of inflow into bioreactor |
| $c_{out}$ | [g l-1] or [mol l-1] | Concentration of outflow of bioreactor |
| $\dot{V}_{in}$ | [g h-1] | Inflow into bioreactor |
| $\dot{V}_{out}$ | [g h-1] | Outflow of bioreactor |
| $\Delta m_i$ | [g] | Change of mass of species $i$ |
| $q_{i,I}$ | [g g-1 h-1] | Specific rate of species $i$ calculated using the integral viable cell density |
| $q_i$ | [g g-1 h-1] | Specific rate of species $i$ |
| $r_i$ | [g l-1 h-1] | Volumetric rate of species $i$ |
| rS | [g h-1] | Substrate uptake rate |
| rx | [g l-1 h-1] | Volumetric biomass formation rate |

| | | |
|---|---|---|
| $\Delta t$ | [h] | Time span used for deriving the rate |
| $\sigma_i^2$ | Squared units of originating signal | Estimated variance of data point $i$ of the originating signal for rate calculation |
| $U_r$ | [g l-1 h-1] | Absolute uncertainty boundary of rate |
| $U_{y_1}$ | units of originating signal | Absolute uncertainty boundary of data point $i$ of the originating signal for rate calculation |
| $u_r$ | [g l-1 h-1] | Simple measurement uncertainty calculated from measurement uncertainties given as standard deviations |
| $V_R$ | [l] | Volume of bioreactor |
| $\gamma_X$ | [g l-1] | Biomass concentration |
| $\gamma_{X_V}$ | [g l-1] | Viable cell concentration |

## Abbreviations

| | |
|---|---|
| BwF[n] | Low pass Butterworth filter with n mHz cut-off frequency |
| CER | Carbon emission rate |
| DWR[n] | Dynamic window rate calculation with SNR of the rate equal to n |
| FWR[n] | Fixed window rate calculation with a window of n data points |
| IVC | Integral of viable cells |
| KaF | Kalman filter |
| NRMSE | Normalized root mean square error |
| PBR | Phase boundary rate calculation |
| SGF[n] | Savitzky Golay smoothing with a window of n data points |
| SGR[n] | Savitzky Golay first derivative with a window of n data points |
| SNR | Signal to noise ratio |
| SNRr | Signal to noise ratio of the rate |

## References

[1]   G. Walsh, Nat. Biotechnol. **2010**, 28 (9), 917–924. DOI:10.1038/nbt0910-917.

[2]   Bioeconomy - Ensuring Smart Green Growth Eur. - Res. - Eur. Comm. **2015**,. http://ec.europa.eu/research/bioeconomy/ (accessed May 25, 2015).

[3]   U.S. Department of Health and Human Services, **2004**,. http://www.fda.gov/downloads/Drugs/Guidances/ucm070305.pdf (accessed June 15, 2015).

[4]   A.S. Rathore, Trends Biotechnol. **2009**, 27 (9), 546–553. DOI:10.1016/j.tibtech.2009.06.006.

[5]   U.S. Department of Health and Human Services, **2011**,. http://www.fda.gov/downloads/Drugs/Guidances/UCM070336.pdf (accessed May 12, 2015).

[6]   A.S. Rathore, R. Johnson, O. Yu, A.O. Kirdar, A. Annamalai, S. Ahuja, et al., BioPharm Int. **2007**, 20 (10),.

[7]   A.O. Kirdar, K.D. Green, A.S. Rathore, Biotechnol. Prog. **2008**, 24 (3), 720–726. DOI:10.1021/bp0704384.

[8]   D. Levisauskas, V. Galvanauskas, S. Henrich, K. Wilhelm, N. Volk, A. Lübbert, Bioprocess Biosyst. Eng. **2003**, 25 (4), 255–262. DOI:10.1007/s00449-002-0305-x.

[9]   C. Dietzsch, O. Spadiut, C. Herwig, Microb. Cell Factories. **2011**, 10 (1), 14. DOI:10.1186/1475-2859-10-14.

[10] P. Sagmeister, T. Langemann, P. Wechselberger, A. Meitz, C. Herwig, Microb. Cell Factories. **2013**, 12 (1), 94.

[11] P. Sagmeister, P. Wechselberger, C. Herwig, PDA J. Pharm. Sci. Technol. **2012**, 66 (6), 526–541. DOI:10.5731/pdajpst.2012.00889.

[12] M. Jenzsch, S. Gnoth, M. Kleinschmidt, R. Simutis, A. Lübbert, J. Biotechnol. **2007**, 128 (4), 858–867. DOI:10.1016/j.jbiotec.2006.12.022.

[13] C. Herwig, U. von Stockar, Bioprocess Biosyst. Eng. **2002**, 24 (6), 395–403. DOI:10.1007/s00449-001-0277-2.

[14] C. Herwig, I. Marison, U. von Stockar, Biotechnol. Bioeng. **2001**, 75 (3), 345–354.

[15] D. Zalai, C. Dietzsch, C. Herwig, O. Spadiut, Biotechnol. Prog. **2012**, 28 (3), 878–886. DOI:10.1002/btpr.1551.

[16] P. Wechselberger, P. Sagmeister, C. Herwig, Biotechnol. Prog. **2013**, 29 (1), 285–296. DOI:10.1002/btpr.1649.

[17] J. Villadsen, J. Nielsen, G. Lidén, Design of Fermentation Processes, in: Bioreact. Eng. Princ., Springer US, **2011**: pp. 383–458. http://link.springer.com/chapter/10.1007/978-1-4419-9688-6_9 (accessed January 12, 2016).

[18] J. Shevit, J. Bonham-Carter, J. Lim, A. Sinclair, BioPharm Int. **2011**, 24 (2),. http://www.biopharminternational.com/economic-comparison-three-cell-culture-techniques (accessed January 11, 2016).

[19] E. Heinzle, A. Oeggerli, B. Dettwiler, Anal. Chim. Acta. **1990**, 238 101–115.

[20] R. Van der Heijden, J.J. Heijnen, C. Hellinga, B. Romein, Kc. Luyben, Biotechnol. Bioeng. **1994**, 43 (1), 3–10.

[21] P. Wechselberger, P. Sagmeister, C. Herwig, Bioprocess Biosyst. Eng. **2012**, 36 (9), 1205–1218. DOI:10.1007/s00449-012-0848-4.

[22] D. Paulsson, R. Gustavsson, C.-F. Mandenius, Sensors. **2014**, 14 (10), 17864–17882. DOI:10.3390/s141017864.

[23] P.R. Patnaik, AIChE J. **2004**, 50 (7), 1640–1646. DOI:10.1002/aic.10156.

[24] J.E. Claes, J.F. Van Impe, Bioprocess Eng. **1999**, 21 (5), 389–395.

[25] M. Arndt, S. Kleist, G. Miksch, K. Friehs, E. Flaschel, J. Trierweiler, et al., Comput. Chem. Eng. **2005**, 29 (5), 1113–1120. DOI:10.1016/j.compchemeng.2004.11.011.

[26] J. Johnston, J. Johnston, J. DiNardo, Econometric Methods, McGraw-Hill, **1996**.

[27] B. Sonnleitner, O. Käppeli, Biotechnol. Bioeng. **1986**, 28 (6), 927–937. DOI:10.1002/bit.260280620.

[28] P. Wechselberger, P. Sagmeister, H. Engelking, T. Schmidt, J. Wenger, C. Herwig, Bioprocess Biosyst. Eng. **2012**, 35 (9), 1637–1649. DOI:10.1007/s00449-012-0754-9.

[29] C. Gmeiner, A. Saadati, D. Maresch, S. Krasteva, M. Frank, F. Altmann, et al., Microb. Cell Factories. **2015**, 14 (1), 1. DOI:10.1186/s12934-014-0183-3.

[30] S. Butterworth, Wirel. Eng. **1930**, 7 (6), 536–541.

[31] T.W. Parks, C.S. Burrus, Digital Filter Design, Wiley-Interscience, New York, NY, USA, **1987**.

[32] S.J. Orfanidis, Introduction to Signal Processing, Prentice Hall, **1996**.

[33] R.E. Kalman, J. Basic Eng. **1960**, 82 (1), 35–45. DOI:10.1115/1.3662552.

[34] P. Wechselberger, C. Herwig, Biotechnol. Prog. **2012**, 28 (1), 265–275. DOI:10.1002/btpr.700.

40

[35] J. Baili, S. Lahouar, M. Hergli, I.L. Al-Qadi, K. Besbes, NDT E Int. **2009**, 42 (8), 696–703. DOI:10.1016/j.ndteint.2009.06.003.

[36] C. Perrin, B. Walczak, D.L. Massart, Anal. Chem. **2001**, 73 (20), 4903–4917. DOI:10.1021/ac010416a.

[37] J.Y. Stein, John Wiley & Sons, Digital signal processing a computer science perspective, Wiley, New York, **2000**. http://search.ebscohost.com/direct.asp?db=iih&jid=%22VLD%22&scope=site (accessed April 23, 2015).

[38] M. Aehle, K. Bork, S. Schaepe, A. Kuprijanov, R. Horstkorte, R. Simutis, et al., Cytotechnology. **2012**, 64 (6), 623–634. DOI:10.1007/s10616-012-9438-1.

[39] R. Van der Heijden, B. Romein, J.J. Heijnen, C. Hellinga, Kc. Luyben, Biotechnol. Bioeng. **1994**, 43 (1), 11–20.

[40] P.W. Sauer, J.E. Burky, M.C. Wesson, H.D. Sternard, L. Qu, Biotechnol. Bioeng. **2000**, 67 (5), 585–597.

# Propagation of Measurement Accuracy to Biomass Soft-Sensor Estimation and Control Quality

Valentin Steinwandter[1*], Thomas Zahel[1*], Patrick Sagmeister[1], Christoph Herwig[2,3,#]

[1] Dipl.-Ing. Valentin Steinwandter, Dipl.-Ing. Thomas Zahel, Dipl.-Ing. Dr.techn. Patrick Sagmeister, Exputec GmbH, Pfeilgasse 32/20, Vienna, Austria

[2] Univ.Prof. Dipl.-Ing. Dr.techn. Christoph Herwig, Institute of Chemical Engineering, Research Area Biochemical Engineering, Vienna University of Technology, Gumpendorferstrasse 1a, Vienna, Austria

[3] Univ.Prof. Dipl.-Ing. Dr.techn. Christoph Herwig, CD Laboratory on Mechanistic and Physiological Methods for Improved Bioprocesses, Vienna University of Technology, Gumpendorferstrasse 1a, Vienna, Austria

[*] These authors contributed equally to this work

[#] Corresponding author. Christoph Herwig, christoph.herwig@tuwien.ac.at, Tel: +43 58801 166400

Abstract

In biopharmaceutical process development and manufacturing the online measurement of biomass and derived specific turnover rates is a central task to physiologically monitor and control the process. However, hard-type sensors such as dielectric spectroscopy, broth fluorescence or permittivity measurement bury various disadvantages. Therefore, soft-sensors, which use measurements of the off-gas stream and substrate feed to reconcile turnover rates and provide an online estimate of the biomass formation, are smart alternatives. For the reconciliation procedure, mass and energy balances are used together with accuracy estimations of measured conversion rates, which were so far arbitrarily chosen and static over the entire process. In this contribution we present a novel strategy within the soft-sensor framework to propagate uncertainties from measurements to conversion rates and demonstrate the benefits: For industrial relevant conditions, hereby the error of the resulting estimated biomass formation rate and specific substrate consumption rate could be decreased by 43 and 64 %, respectively, compared to traditional soft-sensor approaches. Moreover, we present a generic workflow to determine the required raw signal accuracy to obtain predefined accuracies of soft-sensor estimations. Thereby appropriate measurement devices and maintenance intervals can be selected. Furthermore, using this workflow, we demonstrate that the estimation accuracy of the soft-sensor can be additionally and substantially increased.

Keywords

Bioprocess, biomass estimation, soft-sensor, accuracy, error propagation, bioprocess control

## Abbreviations

| | |
|---|---|
| $F_{a,in}$ | Air flow in (L min$^{-1}$) |
| $F_{a,out}$ | Air flow out (L min$^{-1}$) |
| $V_m$ | Molar volume (L mol$^{-1}$) |
| $q_S$ | Specific substrate uptake rate (mol mol$^{-1}$ h$^{-1}$) |
| $r_X$ | Biomass formation rate (mol h$^{-1}$) |
| μ | Specific growth rate (h$^{-1}$) |
| $r_i$ | Consumption/formation rate for species $i$ (mol h$^{-1}$) |
| $S$ | Substrate, C-normalized (mol) |
| $X$ | Biomass, C-normalized (mol) |
| $Y_{X/S}$ | Biomass/substrate yield coefficient |
| $y_{O_2,in}$ | Oxygen fraction in the inlet air (-) |
| $y_{O_2,out}$ | Oxygen fraction in the off-gas stream (-) |
| $y_{CO_2,in}$ | Carbon dioxide fraction in the inlet air (-) |
| $y_{CO_2,out}$ | Carbon dioxide fraction in the off-gas stream (-) |
| $y_{wet}$ | Oxygen fraction in the off-gas without microbial activity (-) |
| $\gamma_i$ | Degree of reduction for species $i$ (-) |
| $\varepsilon_i$ | Applied relative error on species $i$ (-) |
| MFC | Mass flow controller |
| MPD | Median percentage of difference |
| $CER$ | Carbon dioxide evolution rate (mol h$^{-1}$) |
| $OUR$ | Oxygen uptake rate (mol h$^{-1}$) |
| $Ra_{inert}$ | Inert gas ratio (-) |
| $\Delta y$ | Absolute measurement error of signal $y$ |
| $E$ | Elemental composition matrix |
| $\varepsilon$ | Residual vector for non-closing balances |

# 1  Introduction

Biotechnological process development, analysis and control is key to obtain robust processes providing highest product quality attributes as well as a reduced time-to-market latency. Catalyzed by regulatory initiatives for biopharmaceutical products, Process Analytical Technology (PAT) emerged as a major tool that demands for bioprocess analysis and control by frequently measurements ensuring specified final product quality [1]. Especially, in biopharmaceutical production and process development of heterologous protein expression the physiological state of the cells is highly related to the formation of critical quality attributes [2,3]. Therefore, time-resolved knowledge about physiological parameters, such as the specific growth rate or specific substrate uptake rat, is essential in the PAT framework as well as to perform process-development, -characterization and –validation [4]. Moreover, those variables frequently serve as targets for control strategies [5–7]. The key to this physiological information is the catalyst concentration – the biomass. However, the required on-line measurement of biomass is a critical endeavor using hard-type sensors such as dielectric spectroscopy, broth fluorescence or permittivity measurements, each connected to limitations and drawbacks as outlined elsewhere [8]. Software sensors, or short soft-sensors, provide an elegant, non-invasive way to estimate biomass concentration using different other, easy-accessible measurements [9].

In this contribution we want to focus on a dominant biotechnological process mode, the microbial fed-batch fermentation, and on the improvement of one of the most mature soft-sensor implementations using off-gas and substrate-feed measurements. These soft-sensors are established tools for bioprocess control and analysis, which was also frequently shown in practical applications [5,10,11]. Briefly, mass conservation laws are used to calculate turnover rates from online measurements, which might be superimposed with signal errors. In a second step accuracy of turnover rates and constraints, formulated as first-order principles such as mass and energy conservation laws, are used to reconcile the inaccurate turnover rates in order to optimally obey the constraints. Finally the reconciled turnover rates are used to calculate the biomass formation rate ($r_X$), which leads after simple integration over time to the biomass concentration. The resulting information can be used to calculate specific turnover rates, such as the specific substrate uptake rate ($q_S$), which frequently serves as a control variable [12]. Therefore, $r_X$ and $q_S$ are regarded as the most prevailing benchmark entities to evaluate biomass estimation- and physiological control- capability.

However, the control quality by soft-sensors is limited by measurement errors of raw signals used to derive the measured turnover rates. When it comes to industrial applicability, the ultimate question is: Which measurement accuracy is required in order to obtain a sufficiently accurate estimation of the reconciled rates and the biomass?

This question can only be answered if the error sources, their respective impact and possible counteractions are understood. We note that we use the definition of errors as deviations to the true values, excellently defined elsewhere [13]. Random errors leading to a lack of signal precision are caused by small changes within the system, e.g. air movement, temperature and electrostatic fluctuations. A multitude of algorithms exists to smooth signals with random errors ranging from simple median filters to polynomial filters such as Savitzky-Golay filter up to frequency filters such as the Butterworth filter.

While random errors can be minimized quite easily, this is not the case for systematic errors caused by miscalibrations, inaccuracy of analytical devices or a defective feature in the sensor.

Those systematic errors can only be detected and possibly reduced by making use of all available information in terms of first-principle constraints and the accuracy of turnover rates in reconciliation procedures as sketched above. First order principles can be generically formulated for defined processes whereas the accuracy of turnover rates, which are input to the reconciliation procedure, are not known *a priori*. They highly depend on the accuracy of the raw signal measurements and dynamically change over time. It is an existing unmet need to establish a methodology that leverages the accuracy information of the raw signals onto the derived turnover rates, which are subsequently used in the reconciliation procedure.

Therefore, it is goal of this contribution to develop an error propagation procedure to derive the accuracy of turnover rates and demonstrate its benefits in terms of increased physiological accuracy within the soft-sensor framework in microbial fed-batch mode. Moreover, we want to address the question raised above, and present a novel generic workflow that identifies tolerable measurement errors of combinations of multiple analytical measurements in order to meet desired accuracy of soft-sensor estimations.

# 2 Material and methods

## 2.1 Aim and relevance of the presented approach

The following study was carried out with *in silico* generated data. Aim of the *in silico* data generation was to obtain representative microbial fed-batch fermentation data including an induction phase, the predominant industrial mode for the production of recombinant proteins. The experiments were based on an *E. coli* process with oxidative growth and glucose as substrate. As the batch phase is not part of the discussed soft-sensor, only the fed-batch part was considered here.

As commonly used in industry, the modelled fed-batch phase started with an exponential feeding profile. After 8 hours, the induction phase started with a linear feed rate. Starting with the induction phase, the biomass yield coefficient typically decreases during the process and especially fast during the induction phase due to the metabolic load [14]. This can be measured by the soft-sensor and was also considered in the data generation process (Figure 1).

The advantages of an in silico study are obvious:

- It is possible to "run" a bioprocess completely without any errors on the signals and to introduce defined errors into the system. This is not possible with real data, as the exact "real" values without errors on the data cannot be determined.
- A virtually infinite number of experiments with different combinations of errors can be carried out. This enables a systematic study of errors in a high-dimensional "uncertainty space".

## 2.2 Computational environment

All calculations were conducted in a MATLAB environment (2015a, The MathWorks, Inc.). The mechanistic model was created in form of a system of ordinary differential equations. As graphical user interface and bioprocessing toolbox inCyght (2016.02, Exputec GmbH) was used.

## 2.3 In silico data generation

### 2.3.1 Main mechanistic assumptions

The main mechanistic assumptions behind data generation and soft-sensor are the same. Substrate, ammonia and oxygen are converted to biomass and carbon dioxide. In this simple

case, the extracellular formation of product or metabolites will be neglected. This assumption is true for many biopharmaceutical processes, as the product formation rate often is several order of magnitudes lower than the biomass formation rate [15]. For processes were this assumption has to be rejected, the soft-sensor framework has to be extended by on-line product measurement, e.g. by using spectroscopic techniques [16].

$$r_S CH_{pH}O_{pO} + r_{O2}O_2 + r_N NH_3 \rightarrow r_X CH_{zH}O_{zO}N_{zn} + r_{CO2}CO_2$$
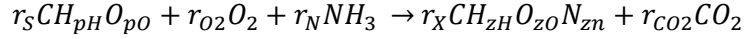
Two first principle assumptions were made; the carbon balance:

$$r_S + r_X + r_{CO_2} = 0$$

And the degree of reduction balance:

$$r_S\gamma_S + r_X\gamma_X + r_{O_2}\gamma_{O_2} = 0$$

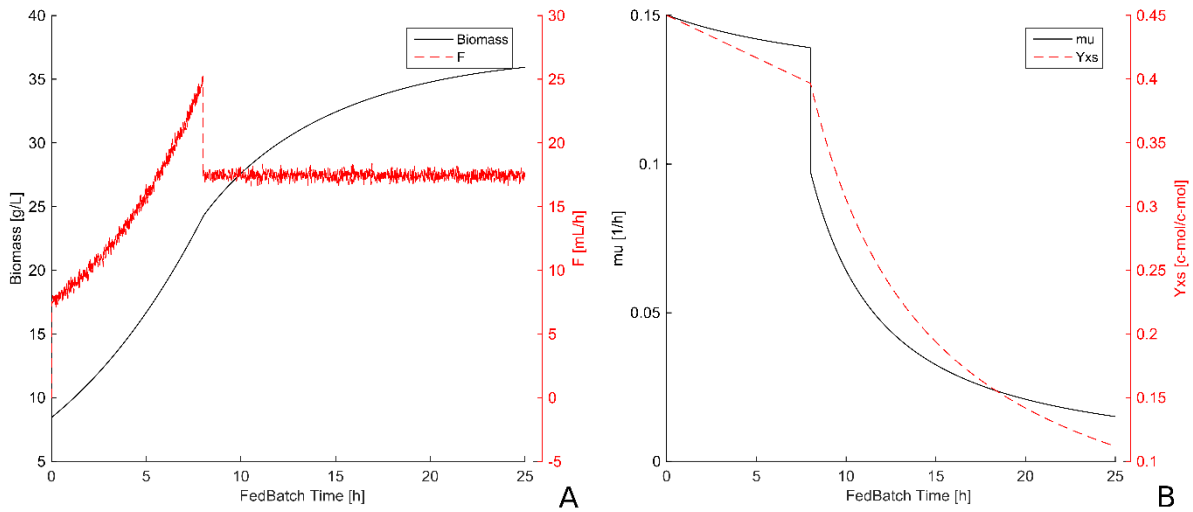The detailed list of equations for the data generation step are shown in the supporting information (section 9.2).



*Figure 1: A: Simulated feed profile and biomass concentration. B: Simulated trajectories of the biomass/substrate yield (Yxs) and the specific growth rate (µ) (right).*

### 2.3.2 Addition of noise and errors on the data

To test the original and new soft-sensors witherroneous data, both systematic as well as random errors were introduced into the model. Based on information of off-gas sensor and mass flow controller manufacturer, as summarized in Table 1, realistic amounts of systematic errors were superimposed to the off-gas data which were used as input for the soft-sensors.

$O_2$ and $CO_2$ concentrations in the exhaust gas are simply applied on the model output for the off-gas data:

$$X_{O_2,measured} = X_{O_2,model} * (1 + \varepsilon_{O_2})$$

$$X_{CO_2,measured} = X_{CO_2,model} * (1 + \varepsilon_{CO_2})$$

As the error on the mass flow controller affects both total oxygen and carbon dioxide input into the system and the resulting final concentrations of $O_2$ and $CO_2$, the error has to be given as

46

input to the model. The set-points for the MFC are the known values, but the model input and real values are calculated as follows.

$$F_{O_2,in,model} = \frac{F_{O_2,in,setpoint}}{1 + \varepsilon_{MFC}}$$

$$F_{CO_2,in,model} = \frac{F_{O_2,in,setpoint}}{1 + \varepsilon_{MFC}}$$

For the errors in the feed rate, a relative error on the set-point rate is applied.

$$r_{S,model} = \frac{r_{S,setpoint}}{1 + \varepsilon_{r_S}}$$

The amounts of systematic errors applied for the different experiments are listed in Section 2.5.

The model delivered an online value each seven seconds. For the addition of random error white Gaussian noise was added to the off-gas signals. The noise was generated by using MATLAB's awgn function with a relative standard deviation of 1 % for the off-gas data and 10 % on the feed rate. The noise on the feed rate is typically relatively high, as the signal often is calculated by deriving the scale signal.

*Table 1: Typical measurement errors of off gas analyzers and mass flow controllers.*

|  | Relative error to measurement value | Measurement accuracy (zero deviance) | Drift/year |
| --- | --- | --- | --- |
| $\Delta F_{a,in}$ (mass flow controller) | ±0.5% of readout | ±0.3-1 % of full scale | ±1 % of full scale |
| $\Delta y_{CO_2,out}$ (infrared) | n.a. | ±1 % of full scale | ±1 % of full scale |
| $\Delta y_{O_2,in}$ (paramagnetic) | ±3 % of readout | ±0.2 % full scale | ±2 % value |
| $\Delta y_{O_2,in}$ (Galvanic cell) | ±3 % of readout | ±0.2 % full scale | ±2 % value |

## 2.4 Quantitative evaluation of bioprocess data and error propagation

### 2.4.1 Preprocessing

As described in the introduction, random errors can be minimized by using preprocessing methods. We decided to apply a Savitzky-Golay filter with a window size of 30 min and 2nd degree polygon on the off-gas signals. These parameters in most cases showed a low signal distortion, while on the other hand the elimination of noise was good. However, it has to be noted that for specific filtering and smoothing problems better filters and filter parameters may exist. In our experience, most of them are not generically applicable, meaning that if they work very well for a specific problem on a defined signal with specific signal dynamics, they may completely fail on another.

### 2.4.2 Data driven rate calculation

The aim of the next section is to express estimators for those conversion rates derived from measurements. In general all conversion rates can be formulated using the simple idea, that the conversion rate equals the net accumulation within the reactor minus the inflow into the reactor plus the outflow out of the reactor.

For demonstration purpose of the subsequent error propagation, the calculation of the conversion rate for $CO_2$ will be calculated exemplarily:

$$r_{CO_2} = CER = \frac{d(CO_2)}{dt} - \dot{CO}_{2,in} + \dot{CO}_{2,out}$$

The term $\frac{d(CO_2)}{dt}$ can be neglected since it is predominantly a function of pH and temperature, which were kept constant over all in silico simulations. Therefore the carbon emission rate ($CER$) formulates to:

$$CER = \frac{F_{a,in}}{V_m}(y_{CO_2,out} \cdot Ra_{inert} - y_{CO_2,in})$$

Where $Ra_{inert}$ is the inert gas ratio, which connects the inflow to the outflow by:

$$Ra_{inert} = \frac{F_{a,out}}{F_{a,in}}$$

And is defined as:

$$Ra_{inert} = \frac{1 - y_{O_2,in} - y_{CO_2,in}}{y_{O_2,out} - y_{CO_2,out} - \frac{y_{wet}}{y_{O_2,in}}}$$

Here $y_{wet}$ is the oxygen concentration in the off-gas stream without bio-reaction and indirectly relates to water stripping out of the reactor.

Well known procedures can be applied in order to calculate the substrate- and oxygen- uptake rate for a substrate limited *E.coli* fermentation as described elsewhere [15].

### 2.4.3  Error propagation

In general, all the input signals for estimating the conversion rates are random variables, associated with a random and systematic error, therefore the estimators itself are random variables, too. As discussed in the Section 1, random errors in the raw signals can be minimized using pre-processing methods whereas systematic errors cannot be removed and propagate directly to the estimated conversion rates. However, via Gaussian error propagation it is possible to estimate the expected error of the conversion rates. This knowledge will subsequently help us to formulate a much more robust reconciliation procedure and estimation of biomass.

The influence of the absolute measurement error ($\Delta y$) of the signal $y$ onto a derived signal $r$ can be approximated using a Taylor expansion [3]:

$$r(y + \Delta y) = r(y) + \frac{1}{1!}\frac{dr(y)}{dy} \cdot \Delta y + \frac{1}{2!}\frac{d^2r(y)}{dy^2} \cdot (\Delta y)^2 + \cdots$$

We want to note that the absolute measurement error $\Delta y$ of the measurement signal can be in most cases calculated from technical device data sheets given by their maximal amplitude (e.g. ±3 % of readout). Therefore, the absolute measurement error $\Delta y$ can be seen as worst-case error. For an approximate solution the Taylor expansion can be terminated after the second term and the resulting absolute deviation of the derived signal ($\Delta r$) can be written as:

$$r(y + \Delta y) - r(y) = \Delta r = \frac{dr(y)}{dy} \cdot \Delta y$$

If the derived signal (here the conversion rate) depends on more than one input variable and the error of the input signal is only known by its boundaries, which is the typical case for biotechnological applications, we can write in analogy:

$$\Delta r = \left|\frac{\partial r}{\partial y_1}\right| \cdot \Delta y_1 + \left|\frac{\partial r}{\partial y_2}\right| \cdot \Delta y_2 + \cdots$$

For the $CER$ the error propagation this formulates to:

$$\Delta CER = \left|\frac{\partial CER}{\partial F_{a,in}}\right| \cdot \Delta F_{a,in} + \left|\frac{\partial CER}{\partial V_m}\right| \cdot \Delta V_m + \left|\frac{\partial CER}{\partial y_{CO_2,out}}\right| \cdot \Delta y_{CO_2,out} + \left|\frac{\partial CER}{\partial Ra_{inert}}\right| \cdot \Delta Ra_{inert}$$
$$+ \left|\frac{\partial CER}{\partial y_{CO_2,in}}\right| \cdot \Delta y_{CO_2,in}$$

$$\Delta CER = \left|\frac{1}{V_m}\left(y_{CO_2,out} \cdot Ra_{inert} - y_{CO_2,in}\right)\right| \cdot \Delta F_{a,in} + \left|\frac{F_{a,in}}{V_m^2}\left(y_{CO_2,out} \cdot Ra_{inert} - y_{CO_2,in}\right)\right| \cdot \Delta V_m$$
$$+ \left|\frac{F_{a,in}}{V_m} \cdot Ra_{inert}\right| \cdot \Delta y_{CO_2,out} + \left|\frac{F_{a,in}}{V_m} \cdot y_{CO_2,out}\right| \cdot \Delta Ra_{inert} + \left|\frac{F_{a,in}}{V_m}\right| \cdot \Delta y_{CO_2,in}$$

$$\Delta R_{a,inert} = \left|\frac{1 - y_{O_2,in} - y_{CO_2,in}}{\left(y_{O_2,out} - y_{CO_2,out} - \frac{y_{wet}}{y_{O_2,in}}\right)^2}\right| \cdot \Delta y_{O_2,out} + \left|\frac{1 - y_{O_2,in} - y_{CO_2,in}}{\left(y_{O_2,out} - y_{CO_2,out} - \frac{y_{wet}}{y_{O_2,in}}\right)^2}\right| \cdot \Delta y_{CO_2,out}$$
$$+ \left|\frac{1 - y_{O_2,in} - y_{CO_2,in}}{\frac{1}{y_{O_2,in}} \cdot \left(y_{O_2,out} - y_{CO_2,out} - \frac{y_{wet}}{y_{O_2,in}}\right)^2}\right| \cdot \Delta y_{wet}$$

This procedure can easily be extended to the $OUR$ and the substrate uptake rate ($r_S$). Typical results of error propagation to the off-gas rates are shown in Figure 2A.

For the presented in-silico study input signals: $F_{a,in}$, $y_{CO_2,out}$ for $CER$ and $y_{O_2,out}$ for the $OUR$, were regarded as superimposed with considerably relevant systematic measurement error. All other input signals were considered to be perfectly accurate. Typical errors for mass flow controllers and off-gas analytics are given in Table 1. However, the error propagation model could be easily extended to more inputs with systematic error. For the error propagation of $r_S$ the only considerably source of systematic signal error was the concentration of the substrate, which might vary due to evaporation during sterilization procedures.
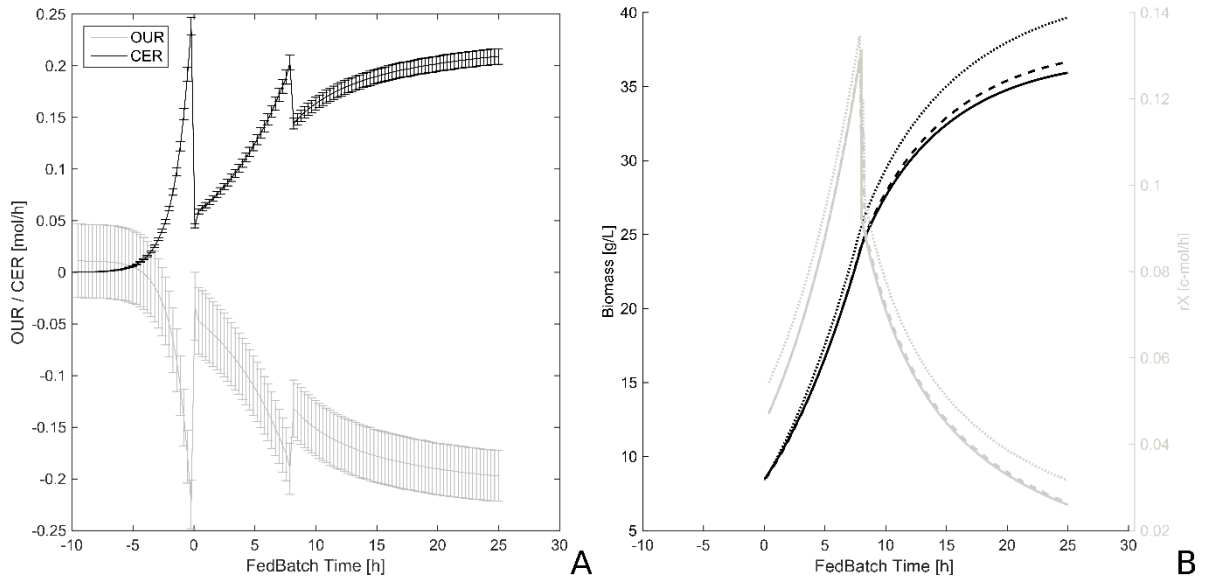
### 2.4.4  Minimum variance rate reconciliation and biomass soft-sensor estimation

In the following section we want to briefly summarize an established minimum variance reconciliation and biomass estimation procedure in order to reduce systematic error on measured turnover rates ($OUR$, $CER$, $r_S$) using first principles as reported in detail elsewhere [15,17,18].

First principles, such as elemental balances (see Section 2.3.1), can be seen as constraints to the bioreactor system. We can formulate many of those constraints and thereby connect components with each other. Commonly a compact matrix formulation is used to connect conversion rates of components with each other using multiple constraints:

$$E \cdot r = 0$$

$E$ is the elemental composition matrix [e x n] with e being the number of elemental balances and $n$ the number of relevant components. $r$ is the vector containing the turnover rates. Under real conditions the elemental balances do not close due to systematic errors of the rates with a residual vector $\varepsilon$:

$$E \cdot r = \varepsilon$$

For this *in silico* example the elemental C balance as well as the degree of reduction (DoR) balance were used as frequently applied elsewhere [5].

In order to identify gross errors in the system it is necessary to check if the residual vector ($\varepsilon$) differs significantly from zero. Therefore, Reilly and Carpani introduced a statistical measure (h-value) which weights the residuals by their accuracy (covariance matrix). Using this $\chi^2$ distributed measure it is possible to set confidence levels for detecting gross errors [18].

The presented concept can easily extended to more elemental balances or energy balances, however, the implementation with C- and DoR-balance is predominantly implemented in industry since additional measurements (e.g. nitrogen or heat transfer) are practically more complex.

The redundancy of the measurable rates (rank of redundancy matrix R [15,17,18]) equals 1 and therefore this redundancy can be used to balance the measured rates in a minimum variance sense and obtain reconciled rates [15].

After no gross error could significantly be detected and the measured rates are reconciled, those reconciled rates can be finally used to estimate the biomass formation rate, which is the only non-measured rate in this example. For this minimum variance balancing procedure the covariance matrix of the measured signals is required. A fair assumption is to state that the covariance of the measured rates is diagonal, which assumes non-correlated errors in the measured signals. In current soft-sensor implementations an empirical approach was chosen and the covariance of all measured rates was assumed to equal 3 % of the readout [10,15]. As a unique feature of the presented soft-sensor implementation, we will use at this point the

derived error boundaries from above as worst case estimators for the variances of the signals. Since the herein derived error boundaries vary dynamically over time, the new approach will be further on called adaptive soft-sensor.

## 2.5 Comparison of soft-sensor estimates to unbiased model data

As a methodology to investigate the result of the soft-sensor as a function of the error of the input signals we investigated 5915 *in silico* experiments with systematically varied errors on the off-gas measurements, substrate concentration and mass flow controlled (described in section 2.3.2). The selected ranges in Table 2 were based on technical manufacturer information of MFC and off-gas analyzer (see Table 1).

As a final output of the soft-sensor the estimated biomass formation rate ($\hat{r}_X$) was compared to the true, unbiased biomass formation rate from the *in silico* model ($r_{X,true}$), which is known. This comparison was done by calculating the Median Percentage of Difference (MPD) over all data points of the time series according to:

$$MPD = 100 * median\left(\frac{\hat{r}_X - r_{X,true}}{r_{X,true}}\right)$$

For each of the 5915 simulations a MPD value for $r_X$ and $q_S$ was calculated. Those values are displayed as surface plots as shown in Figure 3 and Figure 4.

*Table 2: All errors listed here were combined with each other and applied in 5915 experiments. All those in silico generated data sets were used to test the prediction accuracy of traditional and adaptive soft-sensor approach.*

|  | Applied relative error on in- and outputs | Step size |
|---|---|---|
| $\varepsilon_{CO_2},\ \varepsilon_{CO_2}$ | -3 to +3 % | 0.5 % |
| $\varepsilon_{r_S}$ | -3 to +3 % | 1 % |
| $\varepsilon_{MFC}$ | -2 to +2 % | 1 % |

# 3 Results

## 3.1 Comparison of the soft-sensors accuracy

### 3.1.1 Biomass formation rate

In the following sections, a comparison between the traditional approach and the adaptive soft-sensor is done. While traditionally, the errors on $CER$ , $OUR$ and $r_S$ were estimated to be 3 % for all rates and over the whole process, the adapted soft-sensor calculates the accuracy on the rates through error propagation, by making use of the known uncertainty ranges of the raw signals. The herein dynamically resolved accuracy for $OUR$ and $CER$, assuming 3 % maximal error on the read out of offgas analytical measurements of $O_2$ and $CO_2$, are depicted in Figure 2-A. The accurcay of $CER$ is much higher than the accuracy of $OUR$. This information is used by the adapted soft-sensor, therefore, we obtain much more accurate biomass and $r_X$ estimates than previous implementations without error propagation, compared to unbiased biomass and $r_X$ signals, as shown in Figure 2-B.

Since Figure 2B shows only the results for one particular error combination of errors on $O_2$, $CO_2$, $r_S$ and the MFC, we have to resolve the predictions with all other error combinations in order to show superiority of the adapted soft-sensor. Figure 3 shows a comparison of the

biomass formation rate ($r_X$) between the true rates (model) and the estimated rates (left column: traditional soft-sensor, right column: adaptive soft-sensor) by means of MPD. In each of the subfigures the MPD is shown as a function of the error on the $O_2$ and $CO_2$ off-gas concentration, varied between -3 and +3 %. The error on the MFC and on $r_S$ was varied across the rows of the subfigures.

When using the traditional approach (subplots on the left side), especially errors on the oxygen signal lead to high errors on the estimated rates as well as much higher MPD values are reached (up to 42 % compared to maximal 19 % for the adaptive soft-sensor).

In the subplots A1 and A2 in Figure 3 the MPD values regarding $r_X$ are shown as a function solely of error on $O_2$ and $CO_2$ off-gas measurement. In the subplots B1 and B2 we added a relative error of 2 % on $r_S$ to the true model values and in subplots C1 and C2 we added an error of 2 % on the MFC set-point. The average MPD values as well as the maximal MPD values (up to 40 % vs. 18 %) reached throughout all subplots are much lower for the adaptive soft-sensor compared to the traditional approach.
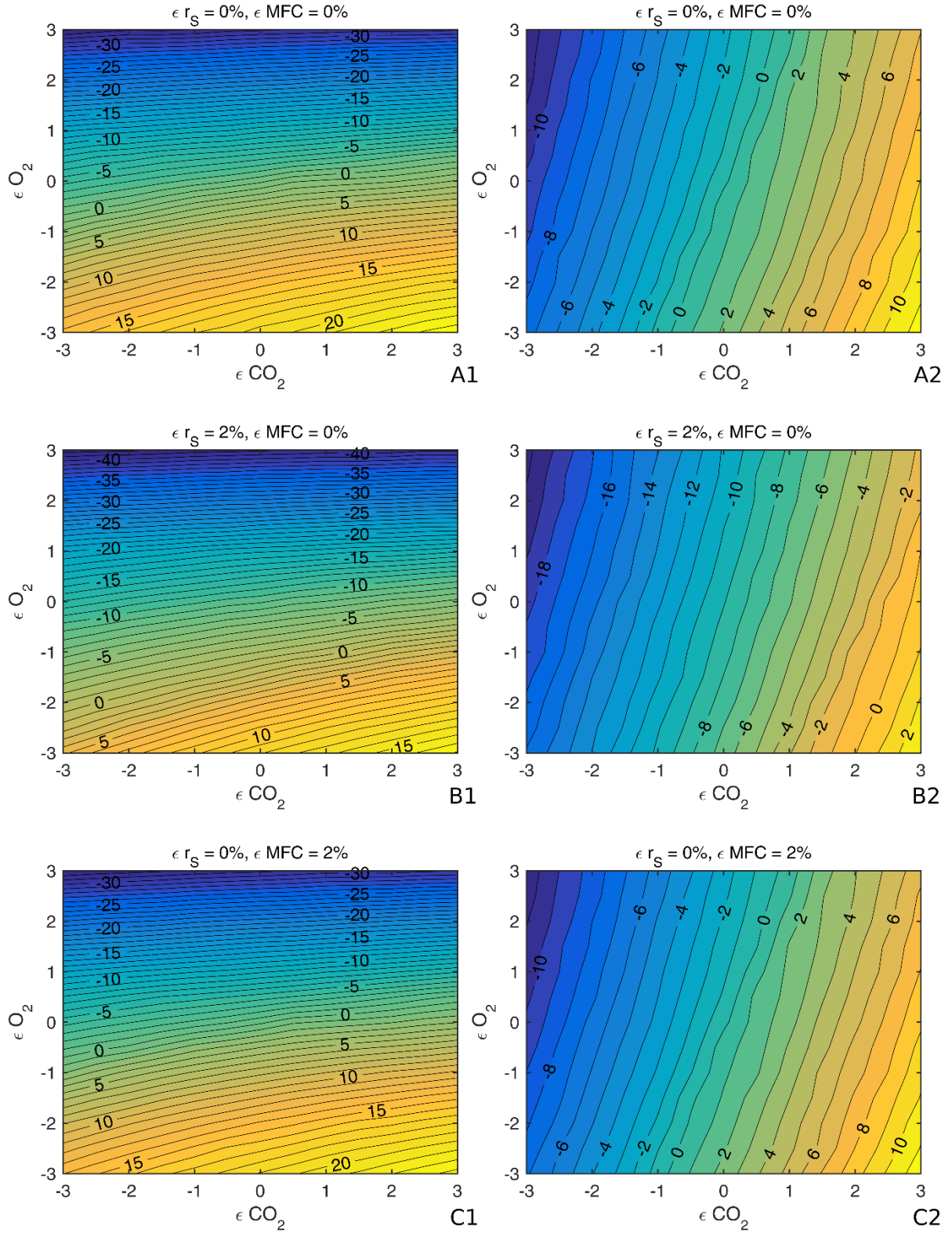
*Figure 3: Comparison of $r_X$ for traditional (left) and adapted (right) soft-sensor, showing the deviation between the real biomass formation rate and the soft-sensor values (%) as a function of the errors on the off-gas data. A: Errors on the off-gas data, but no errors on $r_S$ and the MFC. B: Errors on the off-gas data and 2 % error on $r_S$. C: Errors on the off-gas data and 2 % error on the MFC.*

### 3.1.2 Control quality specific substrate uptake rate

As one of the main use cases of the soft-sensor is the process control based on physiological parameters, the two versions of the soft-sensors were also compared in terms of prediction accuracy for the specific substrate uptake rate $q_S$.

Here, subplots A1 and A2 in Figure 4 show the MPD for $q_S$ with varying error on $O_2$ and $CO_2$ measurement and no error on $r_S$ and MFC signal. Again, the adaptive soft-sensor shows on average much lower MPD values as well as much lower maximal MPD values (up to 15 % for the traditional and up to 4 % for the adaptive soft-sensor). B1 and B2 show, that the estimated relative standard deviation for both soft-sensors is in the area of 3 %. However, when looking at the results in A1 and A2, only the adaptive soft-sensor delivers the estimated standard deviations, since MPD values are in the range of ±3 %.

As shown in C1 and C2 of Figure 4, the h-values of the traditional soft-sensor quickly exceed levels of 3. In this case, the null hypothesis, that there is no gross error in the system, has to be rejected with a confidence level of 95 %. However, the system had no gross error in reality, and as the h-values of the adaptive soft-sensor show, the null hypothesis cannot be rejected when using the correctly calculated covariance matrix for the minimum variance reconciliation. Therefore, the h-values of the traditional soft-sensor have no statistical significance as the covariance matrix, as explained before, is not correctly estimated.
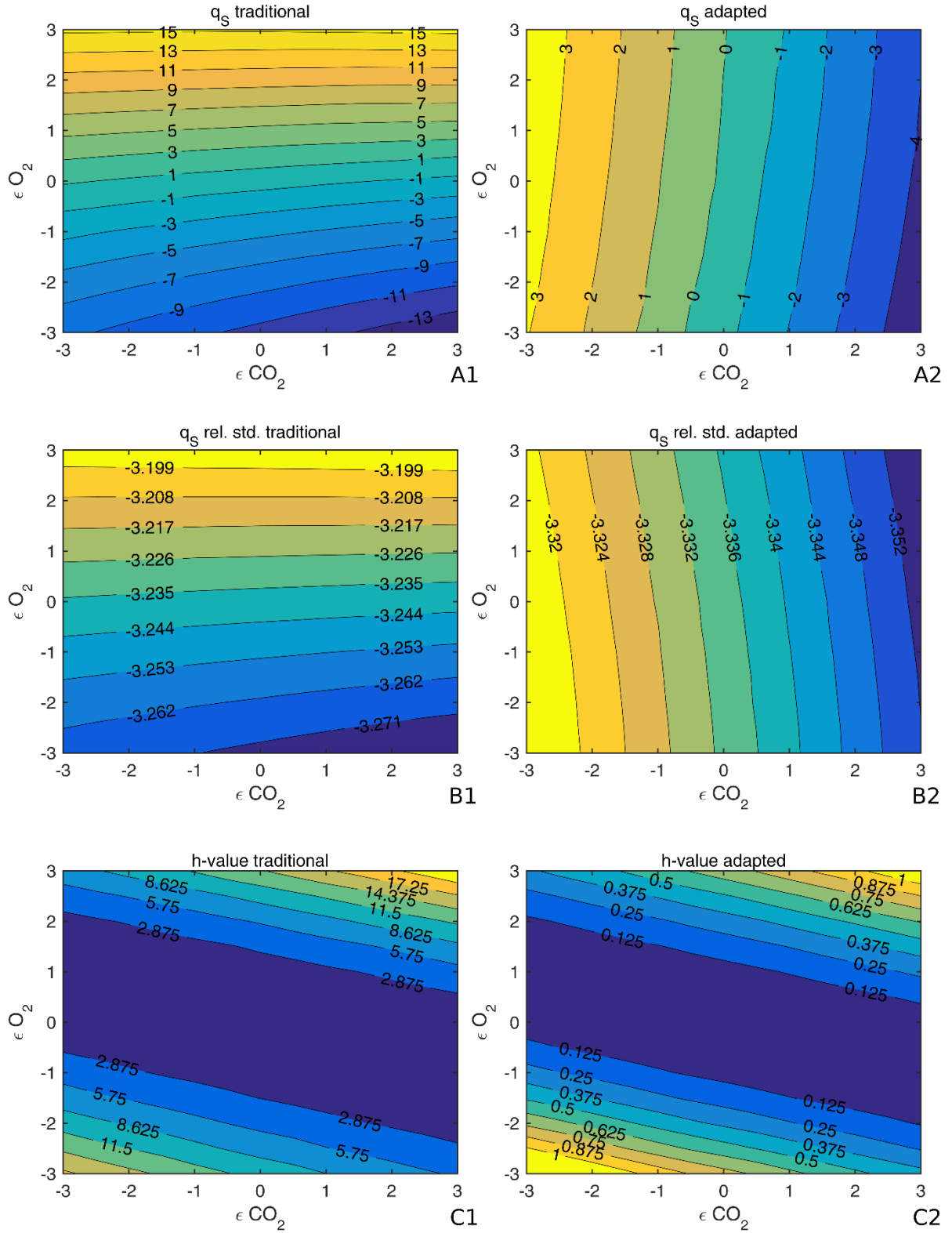
*Figure 4: Comparison of $q_S$ (A), the median of the estimated relative standard deviation on the reconciled $q_s$ (B), and the median h-values (C) for traditional (left) and adapted (right) soft-sensor, depending on the error level of the off-gas analyzers.*

### 3.1.3  Integrated comparison over the entire uncertainty space

The goal of this section is to derive a global parameter which we can use to judge which soft-sensor approach leads to generally more accurate estimations. In general, we face a four dimensional input space consisting of different errors on the $O_2$, $CO_2$, $r_S$ and MFC measurement. This space will be subsequently called the uncertainty space. At each point in this uncertainty space the MPD of the true model value of $r_X$ and $q_S$ is compared to the two soft-sensor approaches. Taking the mean of all those MPD values of the uncertainty space for each soft-sensor approach gives us a clear measure which soft-sensor implementation is generally more accurate. This integrated parameter will be called the global average MPD.

Table 3 summarizes the results of this kind of analysis and shows for each cell the mean MPD of simulations where the $O_2$ and $CO_2$ error was varied between -3 and +3 %, analogous to one subplot of Figure 3. In the columns of Table 2 the error on the MFC is varied, in the rows the error on $r_S$. The global average MPD regarding $r_X$ of the adaptive soft-sensor is 8.7 % compared to 15.2 % of the traditional approach. This is a reduction of the MPD by 43 %. For the estimation of the specific substrate uptake rate $q_S$ the global average MPD can be even lowered from 7.6 to 2.7 which corresponds to a MPD reduction of 64 %.

*Table 3: Comparison of traditional and adaptive soft-sensor for different error levels. Each of the cells show the mean MPD value of simulations in which the error on $O_2$ and $CO_2$ was varied between -3 and +3 %.*

|  | Traditional approach | | | | | Adaptive soft-sensor | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
|  | Mean MPDs of $r_X$ | | | | | | | | | |
|  | $\varepsilon_{MFC}$ [%] | | | | | | | | | |
|  | -2 | -1 | 0 | 1 | 2 | -2 | -1 | 0 | 1 | 2 |
| -3 | 16,1 | 15,9 | 15,7 | 15,6 | 15,4 | 12,7 | 12,7 | 12,7 | 12,7 | 12,7 |
| -2 | 14,8 | 14,6 | 14,4 | 14,3 | 14,1 | 8,6 | 8,6 | 8,6 | 8,6 | 8,6 |
| -1 | 14,3 | 14,1 | 13,9 | 13,7 | 13,5 | 6,0 | 6,0 | 6,0 | 6,0 | 6,0 |
| 0 | 14,4 | 14,2 | 14,0 | 13,8 | 13,6 | 5,2 | 5,2 | 5,2 | 5,2 | 5,2 |
| 1 | 15,0 | 14,8 | 14,6 | 14,4 | 14,2 | 6,2 | 6,2 | 6,2 | 6,2 | 6,2 |
| 2 | 16,1 | 15,9 | 15,8 | 15,6 | 15,4 | 9,1 | 9,1 | 9,1 | 9,1 | 9,1 |
| 3 | 18,0 | 17,9 | 17,7 | 17,6 | <u>17,5</u> | 13,2 | 13,2 | 13,2 | 13,2 | <u>13,2</u> |
|  | Global average MPD of $r_X$: | | | 15,2 | | Global average MPD of $r_X$: | | | 8,7 | |
|  | Mean MPDs of $q_S$ | | | | | | | | | |
| -3 | 8,7 | 8,6 | 8,5 | 8,4 | 8,3 | 3,9 | 3,9 | 3,9 | 3,9 | 3,9 |
| -2 | 8,3 | 8,2 | 8,1 | 8,0 | 7,9 | 3,0 | 3,0 | 3,0 | 3,0 | 3,0 |
| -1 | 7,9 | 7,8 | 7,7 | 7,6 | 7,5 | 2,4 | 2,4 | 2,4 | 2,4 | 2,4 |
| 0 | 7,5 | 7,4 | 7,4 | 7,3 | 7,2 | 2,1 | 2,1 | 2,1 | 2,1 | 2,1 |
| 1 | 7,3 | 7,2 | 7,1 | 7,1 | 7,0 | 2,1 | 2,1 | 2,1 | 2,1 | 2,1 |
| 2 | 7,2 | 7,1 | 7,0 | 7,0 | 6,9 | 2,4 | 2,4 | 2,4 | 2,4 | 2,4 |
| 3 | 7,2 | 7,1 | 7,0 | 6,9 | <u>6,9</u> | 2,8 | 2,8 | 2,8 | 2,8 | <u>2,8</u> |
|  | Global average MPD of $q_S$: | | | 7,6 | | Global average MPD of $q_S$: | | | 2,7 | |

($\varepsilon_{rs}$ [%] is the row axis label)

## 3.2  Generic workflow to ensure appropriate control quality

Besides showing superior behavior of the new soft-sensor implementation over state-of the art, we want to present a novel generic workflow to obtain a desired soft-sensor estimate or

reconciliation quality by adapting accuracy of measurement devices. This workflow includes the following steps as indicated in Figure 5:

1.  Use a mechanistic process model to generate time resolved data which will be used to derive rates. These are used as input to the soft-sensor (here $O_2$ and $CO_2$ concentration of the off-gas stream, substrate concentration and inflowing air controlled by MFC).
2.  Obtain biased signals by superimposing them with representative systematic (see manufacturer specifications) and random noise (estimated process noise).
3.  Calculate turnover rates including their accuracy as described in Section 2.4.2. The rates and their accuracy (covariance) are input to the soft-sensor.
4.  Use the soft-sensor's first principles to reconcile measured turnover rates unless gross errors are detected. The reconciled rates can be used to estimate the biomass and all related entities (e.g. $q_S$ or biomass).
5.  The herein obtained $q_S$ estimate (or biomass estimate) is compared to the true, unbiased model signal. If the estimate does not meet the predefined thresholds (e.g. 5 % global average MPD), more accurate measurement devices and their respective measurement errors are used to continue with step 2 to 5 with reduced systematic error levels. The selection of appropriate measurement devices is driven by technical, manufacturing and financial constraints, which is not scope of this study.
6.  If the estimate meets the predefined thresholds in terms of global average MPD, a robust estimate under industrial relevant process conditions is achieved.

As an example, the error ranges of Table 2 were taken as a starting position in Step 2 of the generic workflow presented in Figure 5. It was assumed that the desired control quality could not be reached with the current analytical devices (Step 5), therefore, exemplary a higher accuracy of the oxygen sensor and MFC from ±3 to ±0.5 % and ±2 to ±1 %, respectively, was implemented. The results before and after this change are shown in Figure 6. After the change, the estimated error surface of $q_S$ is rotated in a favorable direction to enlarge regions of lower error (0 to 2% error), as depicted in the non-shaded areas of the two subfigures of Figure 6. Overall this results in 10 % reduced global average MPD.
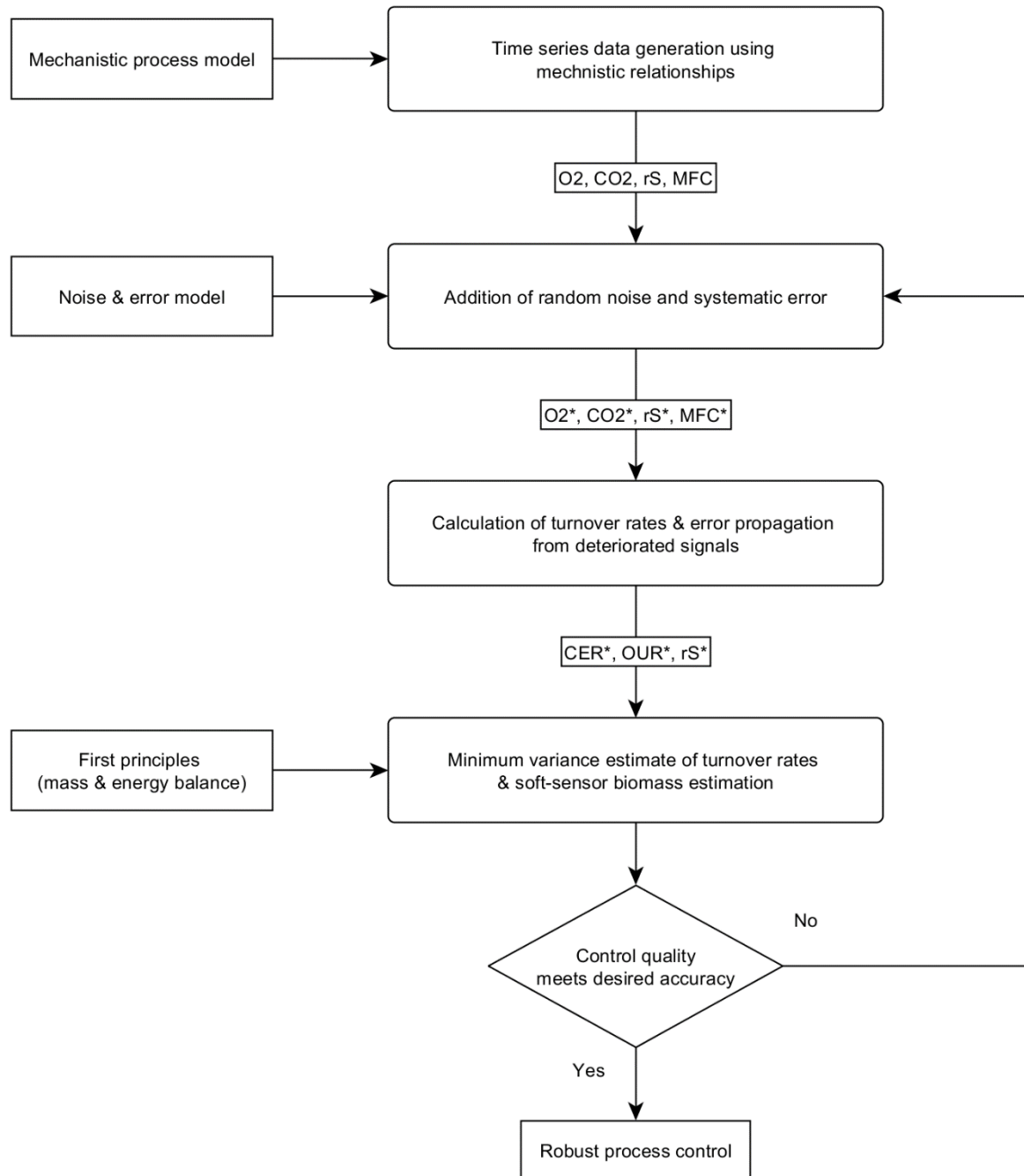
*Figure 5: Generic workflow for identification of desired measurement error and noise for robust biomass soft-sensor estimation. Asterisk indicate variables which were superimposed by random noise and systematic error.*
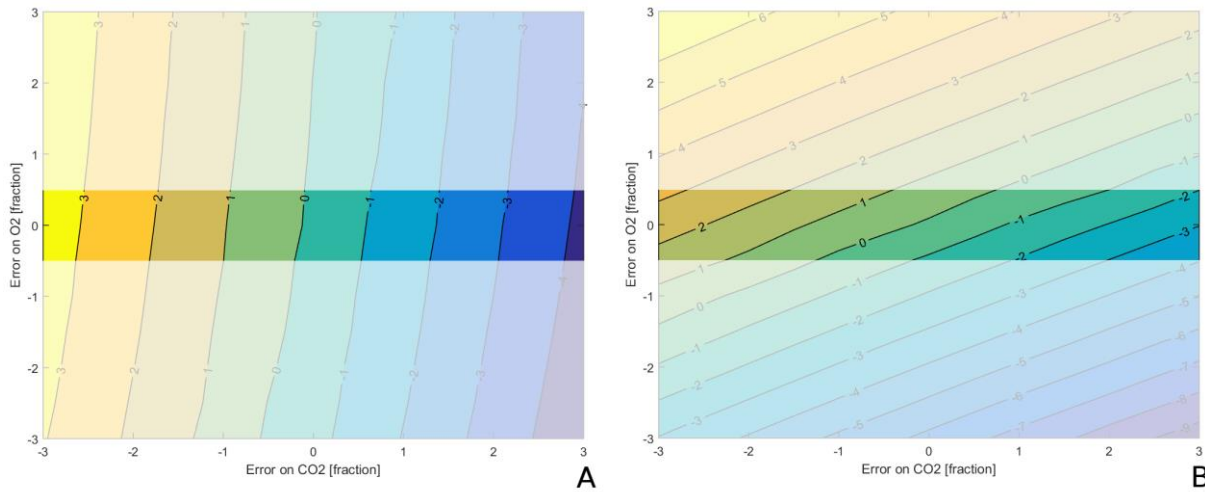
*Figure 6: Estimation error of $q_S$ before (A: ±3 % maximal error on oxygen measurement and ±2% maximal error on MFC) and after (B: ±0.5 % maximal error on oxygen measurement and ±1 % maximal error on MFC) the increase of the input signal accuracies. Not only values outside the uncertainty range can be excluded but also the mean MPD inside the uncertainty space is more accurate.*

# 4 Discussion

## 4.1 Superior accuracy for the estimated rates of the adaptive soft-sensor

As perceived in Figure 2, Figure 3, Figure 4-A1 and Figure 4-A2 as well as summarized in Table 1, the adaptive soft-sensor delivers more accurate estimates of $r_X$, which integrates to biomass, and $q_S$. Moreover, maximal MPD values for $r_X$ and $q_S$ are much lower for the adaptive soft-sensor which implies that the biomass estimate as well as the control of $q_S$ can be performed much more robust under real process conditions since large deviations to the true values of $r_X$ and $q_S$ can be avoided.

This is due to the fact that the covariance matrix for the minimum variance reconciliation procedure is arbitrary chosen for the traditional soft-sensor, which assumes a too low uncertainty range for the $OUR$, as shown exemplarily in Figure 2. The adaptive soft-sensor on the other hand, dynamically uses all available information (uncertainty ranges of off-gas analysis) to calculate a realistic covariance matrix. This leads to a much more robust estimate of $r_X$ and $q_S$.

For each subfigure there are some "sweet spots" for certain error combinations, where the classical soft-sensor shows a better accuracy in the prediction of $r_X$. Since the exact combination of the present errors on the input signals is not known *a priori*, this is no advantage under real process conditions.

## 4.2 Statistical meaningfulness of standard deviation and h-value

The covariance matrix is a critical input for the minimum variance reconciliation procedure. As for the traditional soft-sensor, the covariance matrix consists of arbitrary values which do not represent the true and dynamically changing uncertainties, all calculated statistical measures, i.e. standard deviation and h-value, lose their significance. This is different for the adaptive soft-sensor.

As shown in Figure 4-B1 and Figure 4-B2, the relative standard deviations of the estimated rates are in the range of 3 % and almost identical for traditional and adaptive soft-sensor.

59

However, when comparing these standard deviations to the actually measured errors in terms of MPD in Figure 4-A1 and Figure 4-A2, it becomes clear that the calculated standard deviations do not fit to these errors for the traditional soft-sensor; the measured MPDs go up to 15 % in the considered area. For the adaptive soft-sensor the standard deviations are meaningful and on the same magnitude as the actually measured MPDs. Under real process conditions the calculated standard deviation of an estimated rate is the only available measure to evaluate their prediction accuracy and expected uncertainty and is therefore of critical importance.

As shown in C1 and C2 of Figure 4, the h-values of the traditional soft-sensor quickly exceed levels of 3. As already explained in Section 3.1.2, the calculated h-values are statistically not meaningful. This means that they cannot be used to detect a gross error in the system with a defined level of significance. They only can be used to relatively compare similar processes or detect gross errors when the h-values are magnitudes higher than expected. This is not true for the adaptive soft-sensor, as over the whole uncertainty space the h-values are below 3 and no false positive detection of gross errors occurred with 95 % confidence.

## 4.3 Applicability of the generic workflow to set measurement accuracies and ensure desired accuracy of soft-sensor estimations

The question about the required measurement accuracy of raw signals to meet the desired accuracy of derived variables, such as the soft-sensor estimation for bioprocess control, is equally urged by device manufacturer as by process engineers. This is due to the fact that measurement accuracy is often correlated to higher asset costs of advanced devices or more frequent maintenance intervals of existing devices.

In Section 3.2 we present a generic workflow to answer this question. Since the measurement accuracy of the derived soft-sensor estimate is not only a function of the accuracy of the input signals (Step 2 of the workflow) but also of the dynamics of the process, a mechanistic model has to provide this information (Step 1 of the workflow). If one has multiple possibilities of exchanging devices or maintenance intervals to increase accuracy of input signals, this can be solved iteratively in the workflow by testing different of those combinations and evaluating if the resulting accuracy of the soft-sensor is sufficient. Moreover, as shown in Figure 6, it is thereby possible to not only get rid of areas with high levels of MPD but rather additionally increase the accuracy in the reduced uncertainty space due to the introduction of supplementary knowledge about the accuracy of input signals.

## 4.4 Extrapolations of the adaptive soft-sensor and the generic workflow to other application areas

The presented error propagation approach as well as the presented workflow are generically applicable to include additional sources of information. For example, the consideration of energy balances based on the metabolic heat production during a process [19] or the already mentioned combination of the soft-sensor with spectroscopic techniques [16] could be included. This would result in an even more robust and diverse applicable package.

# 5 Conclusion

In this contribution we aim to present an error propagation procedure increasing the accuracy and robustness of the soft-sensor estimates.

Traditionally the uncertainties for conversion rates ($CER$, $OUR$, $r_S$) were arbitrarily assumed and static over the whole process. Here we established a novel procedure to obtain meaningful

uncertainties, dynamically changing over time, which are used as representative knowledge source together with first principles in the soft-sensor framework.

In this *in silico* case study the new approach using the adaptive soft-sensor the error on the estimates could be reduced by 43 % for the estimated biomass growth rate ($r_X$) compared to traditional soft-sensor implementations. For the estimation of the specific substrate uptake rate $q_S$ the error on the estimate could even be lowered by 64 %.

When using the traditional soft-sensor approach, the resulting h-values could not be used to statistically reject the null hypothesis of detecting gross errors, since estimations of covariance of the turnover rates were arbitrarily chosen and static over time. The new approach delivers both, statistically meaningful h-values for the detection of gross errors and informative standard deviations on the estimated rates. Latter ones are essential under real process conditions to judge soft-sensor estimation quality, as obviously there exists no possibility to evaluate the control quality by comparing the estimates to unbiased model values.

Additionally we presented a new generic approach to ensure a predefined control quality of the soft-sensor estimate by iteratively evaluating the effect of the different errors on the raw signal measurements. It has been demonstrated that by following this generic workflow it is possible to additionally significantly increase the adaptive soft- sensor accuracy.

The presented approach can be generically applied taking also additional error sources into account. The new methodology is practically applicable to industrial conditions, where maximal errors of measurement devices are used to obtain dynamically changing accuracies of derived turnover rates as shown in Figure 2A.

# 6  Acknowledgement and contributions

# 7  References

1. FDA. Guidance for Industry: PAT - A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance. 2004.
2. Sagmeister P, Wechselberger P, Herwig C. Information Processing: Rate-Based Investigation of Cell Physiological Changes along Design Space Development. PDA J. Pharm. Sci. Technol. 2012;66:526–41.
3. Jenzsch M, Gnoth S, Kleinschmidt M, Simutis R, Lübbert A. Improving the batch-to-batch reproducibility of microbial cultures during recombinant protein production by regulation of the total carbon dioxide production. J. Biotechnol. 2007;128:858–67.

4. Read EK, Shah RB, Riley BS, Park JT, Brorson KA, Rathore AS. Process analytical technology (PAT) for biopharmaceutical products: Part II. Concepts and applications. Biotechnol. Bioeng. 2010;105:285–95.
5. Sagmeister P, Wechselberger P, Jazini M, Meitz A, Langemann T, Herwig C. Soft sensor assisted dynamic bioprocess control: Efficient tools for bioprocess development. Chem. Eng. Sci. 2013;96:190–8.
6. Dabros M, Schuler MM, Marison IW. Simple control of specific growth rate in biotechnological fed-batch processes based on enhanced online measurements of biomass. Bioprocess Biosyst. Eng. 2010;33:1109–18.
7. Riesenberg D, Schulz V, Knorre WA, Pohl H-D, Korz D, Sanders EA, et al. High cell density cultivation of Escherichia coli at controlled specific growth rate. J. Biotechnol. 1991;20:17–27.
8. Kiviharju K, Salonen K, Moilanen U, Eerikäinen T. Biomass measurement online: the performance of in situ measurements and software sensors. J. Ind. Microbiol. Biotechnol. 2008;35:657–65.
9. Luttmann R, Bracewell DG, Cornelissen G, Gernaey KV, Glassey J, Hass VC, et al. Soft sensors in bioprocessing: A status report and recommendations. Biotechnol. J. 2012;7:1040–8.
10. Sagmeister P, Langemann T, Wechselberger P, Meitz A, Herwig C. A dynamic method for the investigation of induced state metabolic capacities as a function of temperature. Microb. Cell Factories 2013;12:94.
11. Jobé AM, Herwig C, Surzyn M, Walker B, Marison I, von Stockar U. Generally applicable fed-batch culture concept based on the detection of metabolic state by on-line balancing. Biotechnol. Bioeng. 2003;82:627–39.
12. Wechselberger P, Sagmeister P, Engelking H, Schmidt T, Wenger J, Herwig C. Efficient feeding profile optimization for recombinant protein production using physiological information. Bioprocess Biosyst. Eng. 2012;35:1637–49.
13. JCGM. Evaluation of measurement data — Guide to the expression of uncertainty in measurement.
14. Wechselberger P, Sagmeister P, Herwig C. Model-based analysis on the extractability of information from data in dynamic fed-batch experiments. Biotechnol. Prog. 2013;29:285–96.
15. Wechselberger P, Sagmeister P, Herwig C. Real-time estimation of biomass and specific growth rate in physiologically variable recombinant fed-batch processes. Bioprocess Biosyst. Eng. 2012;36:1205–18.
16. Lourenço ND, Lopes JA, Almeida CF, Sarraguça MC, Pinheiro HM. Bioreactor monitoring with spectroscopy and chemometrics: a review. Anal. Bioanal. Chem. 2012;404:1211–37.
17. Van der Heijden R, Heijnen JJ, Hellinga C, Romein B, Luyben Kc. Linear constraint relations in biochemical reaction systems: I. Classification of the calculability and the balanceability of conversion rates. Biotechnol. Bioeng. 1994;43:3–10.
18. Van der Heijden R, Romein B, Heijnen JJ, Hellinga C, Luyben Kc. Linear constraint relations in biochemical reaction systems: II. Diagnosis and estimation of gross errors. Biotechnol. Bioeng. 1994;43:11–20.
19. Paulsson D, Gustavsson R, Mandenius C-F. A Soft Sensor for Bioprocess Control Based on Sequential Filtering of Metabolic Heat Signals. Sensors 2014;14:17864–82.

# Workflow for Criticality Assessment applied in Biopharmaceutical Process Validation Stage 1

**Thomas Zahel[1], Lukas Marschall[1], Sandra Abad[2], Elena Vasilieva[2], Daniel Maurer[2], Eric M. Mueller[3], Patrick Murphy[3], Thomas Natschläger[4], Cécile Brocard[2], Daniela Reinisch[2], Patrick Sagmeister[1] and Christoph Herwig[1*]**

[1]  Exputec GmbH, Mariahilferstraße 147, Vienna, Austria
[2]  Boehringer Ingelheim RCV GmbH & Co KG, Doktor-Boehringer-Gasse 5-11, Vienna, Austria
[3]  Versartis Inc., 4200 Bohannon Drive, Suite 250 Menlo Park, California, USA
[4]  Software Competence Center Hagenberg, Softwarepark 21, 4232 Hagenberg, Austria
*  Correspondence: christoph.herwig@exputec.com; Tel: +43-1-58801-166400

**Abstract:** Identification of critical process parameters that impact product quality is a central task during regulatory requested process validation. Commonly, this is done via design of experiments and identification of parameters significantly impacting product quality (rejection of the null hypothesis that the effect equals 0). However, parameters which show a large uncertainty and might result in reaching an undesirable product quality limit critical to the product, may be missed. This might occur during the evaluation of experiments since residual/un-modelled variance in the experiments is larger than expected a priori. Estimation of such a risk is the task of the presented novel retrospective power analysis permutation test. This is evaluated using a data set for two unit operations established during characterization of a biopharmaceutical process in industry.  The results show that for one unit operation the observed variance in the experiments is much larger than expected a priori, resulting in low power levels for all non-significant parameters. Moreover, we present a workflow how to mitigate the risk associated with overlooked parameter effects. This enables a statistically sound identification of critical process parameters. The developed workflow will substantially support industry in delivering constant product quality, reduce process variance and increase patient safety.

**Keywords**: retrospective power analysis; process characterization study; process validation stage 1; criticality assessment; control strategy; design of experiments

## 1. Introduction

Process validation of pharmaceutical processes aims to demonstrate the capability of the process to constantly deliver high product quality [1,2]. Most of the warning letters connected to process validation are raised due to flaws in stage 1 [3]. The aim of process validation stage 1 is to identify a robust process design that enables the ability to constantly deliver product quality. Therefore, it is key to identify critical process parameters (CPPs) that are likely to create risk to critical quality attributes (CQAs) and set up control strategies for these CQAs. Thereby it is possible to reduce out of specification (OOS) events, recalls, and ultimately risk to the patient. At process validation stage 1, it is the highest priority not to overlook a CPP in the design of the process, which as a consequence might not be controlled properly.

In order to accomplish this goal the following steps are commonly undertaken in industry to characterize process design following a risk based approach:

1.  Risk assessment: to identify potential influential/critical parameters for each unit operation: This is usually performed using tools such as failure mode and effect analysis (FMEA) [4,5]. Ranking of potential criticality is performed using expert knowledge, historical process data, and interdependencies identified in development data.

2.  Scale down model establishment: Due to the costs related to large scale experiments, in biopharmaceutical manufacturing it is necessary to develop appropriate scale down models

(SDMs) that are appropriate to investigate the interdependency between process parameters and quality attributes.

3.  Experimental designs: Design of Experiments are applied to quantify the impact of process parameters (PPs) onto CQAs. Prior to conducting experiments, a-priori power analysis is a good practice to evaluate if an effect that leads to a change in product quality – in the following defined as a critical effect – can be detected by the proposed design setting. Statistical power is defined as the probability that we are able to detect an effect if it is truly there [6]. This is done for a priori analysis by estimating the expected signal to noise ratio, which is thought to occur during the experiments [7]. As a result of this a-priori power analysis, the number of required experiments, the intended screening range, or the design itself might be adjusted. After a sufficient power can be expected, potential influential/critical parameters are purposefully varied within experiments, which is done for each unit operation separately using the previously established SDMs.

4.  Criticality assessment of process parameters by evaluating experimental designs: Identification of significant factors (rejection of the null hypothesis that the effect equals 0) at a desired significance level (typically α<0.05) is performed using Pareto charts and analysis of significance of regression coefficients by means of ANOVA. Misleadingly, this does not imply that for non-significant factors the null hypothesis is true and their effect is zero [8]. Rather it indicates that the uncertainty around these factors in the range examined – often indicated by large confidence intervals around the effect - is large and critical levels cannot be excluded. Commonly, only significant factors that have been observed to impact product quality or process performance are defined as critical or key, respectively. Those which cannot be stated as significantly impacting are stated as non-critical or non-key, respectively.

5.  Definition of control strategy: As a means to ensure all CQAs and quality specifications are met a process control strategy for all critical and key process parameters must be put in place. Moreover, it has to be evaluated whether their mutual worst case setting would lead to acceptable product quality levels. Commonly for biopharmaceutical production, this is accomplished by setting normal operating ranges (NOR) and proven acceptable ranges (PAR).

Although all steps are equally important to design a robust process, we frequently observed that in industry steps 3, 4 and 5 are more difficult to accomplish in practice. The US Food and Drug Administration (FDA) and other agencies are not prescriptive but clearly state that statistics should be used within all stages of process validation [3]. Multiple statistical tools and software for step 3 (a priori power analysis & design of experiments) and step 4 (statistical analysis of significant parameters) exist, however, the approach of those steps as described above has two major drawbacks: (i) after making several assumptions about the expected noise in the a priori power analysis of step 4, those assumptions are not checked for validity after the experiments have been performed. Especially in biopharmaceutical engineering, reproduction and analytical variability from non-validated methods, which might be used during stage 1 of process validation, as well as unexpected non-linear effects (e.g. edge of failure experiments), may lead to increased noise in the conducted design of experiments (DoEs). (ii) Criticality and potential tightening of the NOR is only taken into account for significant parameters. This might not be sufficient since parameters with large uncertainty around the estimated effect - those effects, which might be zero, but might be very large, too - can have severe effects on product quality as well.

The first of the mentioned drawbacks can be tackled by retrospective assessment of the actually received power. Although retrospective power analysis is controversially discussed when using the observed variance and observed effect size, it is an appropriate tool when comparing the observed variance in the experiments to a pre-specified critical effect [6,9]. Frequently, retrospective power is calculated using the observed effect size, which leads to uninformative results [10].

Both issues together might lead to situations where the process shows unexpected variability during routine manufacturing. Therefore, we want to present a workflow for criticality assessment that reduces the risk to overlook critical PPs. This is demonstrated based on a process characterization study of a novel long acting human growth hormone product. Exemplarily for two unit operations, we will address the following topics:

•   Establishment of a methodology that prevents engineers, during process validation, from overlooking critical parameters

- Setting a control strategy for critical and likely overlooked parameters that ensures a robust process design
- A workflow that can be used during stage 1 process validation to assess PP criticality. Applying those guidelines it will be possible to better understand potential process variability and provide an opportunity to reduce process variability, OOS events, and patient risk.

## 2. Methods

In the following sections we describe the biopharmaceutical production process, selection of experimental designs to study the impact of PPs on CQAs (section 2.1), calculation procedures for critical effects (section 2.2), an a priori power analysis approach (section 2.3) applied to assess the ability of the DoE to detect practically relevant (here critical) effects and their statistical evaluation (section 2.4).

### 2.1. Description of process and design of conducted experiments

The workflow for criticality assessment will be presented for two unit operations from a biopharmaceutical manufacturing process producing a recombinant protein. The process consists of an Escherichia coli fermentation, cell lysis, precipitation (PR), clarification (depth filtration), and three subsequent preparative chromatographic columns (CC 1/ CC 2/ CC 3) for purification. Finally, ultrafiltration/diafiltration is performed to adjust product concentration. For the presented case study for criticality assessment, unit operations CC 1 and the precipitation step were exemplarily chosen.

Risk assessment (FMEA) conducted by process experts showed that five and four PPs respectively, had a high risk priority number and need to be studied experimentally in respect to their influence on CQAs for CC1 and PR, respectively (see Table 1 and Table 2). Due to the number of studied PPs for both unit operations, a definitive screening design was chosen [11,12]. Except one parameter (Mixing [Yes/No] for precipitation) all DoE factors are numerically scaled. Small scale experiments were used to conduct DoEs.

### 2.2. Calculation of thresholds for critical effects

We formulate a critical gap (CG) as the difference between the performance at set-point conditions and the threshold for each response:

$$CG = theshold_{USL} - \bar{y}(x_{SP}) \tag{1}$$

where $\bar{y}(x_{SP})$ is the response value (here a specific concentration of an impurity) at set-point condition of manufacturing. Since we do not have lower specification limits for the studied impurities, thet threshold, which must not be surpassed, is derived from the upper specification limit (USL) of drug substance (DS) specifications. The studied unit operations are at an intermediate stage of the process. We therefore, calculate the specification limit for the investigated unit operation by multiplying the final DS specifications times the mean specific clearance factors from manufacturing scale of all unit operations in between. This approach might be refined by including knowledge on increased impurity clearance, e.g. due to spiking studies. Choosing the approach with mean specific clearances might seem conservative, however, it is desirable to reduce risk of overestimating the specific impurity clearance. The specific clearance factors for each unit operation are defined by:

$$Specific\ Clearance = SC = \frac{c_{CQA,load}}{c_{CQA,pool}} \tag{2}$$

where cCQA, load and cCQA, pool are the specific concentrations (mg CQA per mg product) of the respective CQA prior and after the unit operation.

$$threhsold_{USL} = USL * \prod_{u=k}^{U} SC_u \tag{3}$$

where u=k…U is counting the unit operations from the studied kth unit operation until the last unit operation (U) which equals DS.

### 2.3. A priori power analysis

We want to investigate if the residual error during evaluation of experimental designs (DoEs) masks effects to an extent such that they could collectively surpass a critical threshold (e.g. specification limit of a specific CQA concentration) within normal operating ranges (see section 0 for calculation of thresholds). Since we are dealing with a multivariate problem we need to identify how many parameters and to which extent each of those parameters contribute to surpassing of such a critical threshold. From a sparsity assumption, it is unlikely that all effects that can be studied using a certain design (e.g. all main effects and interactions effects) are truly present. Therefore, it is a common assumption applied to many statistical packages to study only power of the total number of main effects[13].

Moreover, in multivariate analysis (p>1) infinite combinations of effects of multiple parameter exists that lead to surpassing of such a critical threshold. E.g., the full effect to surpass the critical threshold might be explained solely by the first parameter ($P_1$) and no effect is present from the residual parameters ($P_r$), or a fraction of the entire effect is explained by P1 (e.g. 10%) and the residual 90% are equally explained by $P_r$. Overall, we are interested in the mean chance to detect any of those combinations. Per default, classical statistical software such as JMP (SAS Institute Inc.) or DesignExpert (Stat-Ease, Inc.) only allow for fixed effect power calculation [10,13]. Here we propose a more general method based on the assumption that the effects are randomly distributed over all parameters. Therefore, we assigned weights to the parameters and varied the fraction/weight of the entire effect that is explained by each parameter gradually between 0.0 to 1.0 (we used a step size of 0.01 in our experiments, i.e. 100 steps) and split the residual effect equally under the residual parameters: $w_i = a$, $w_{j \neq i} = (1 - a)/(p - 1)$, for $a = 0, \dots, 1$ and $i = 1, \dots, p$. Hence all the weights $w_i$ sum up to 1. In total, we obtain $C = p * 100$ combinations of possible effect distributions and the resulting power values. The mean for each parameter of these recorded power values was taken as the power for this experimental design (see step 6 of the a priori workflow present below).

Herein, the following workflow for a priori power analysis can be formulated:

1. Estimate the mean ($\bar{y}_{SP}$) and variance ($\sigma_{SP}$) of the response variable from small scale or pilot scale experiments at set point conditions of manufacturing. We assume that residual error in the model is only due to process- and analytical variance. The latter estimate will be used to calculate the expected sum of squares of the residuals ($\widehat{SS_{res}}$):

$$\widehat{SS_{res}} = (n - 1) * \sigma^2_{SSP} \tag{4}$$

2. For each of the combinations (c) described above we calculate critical effects for each parameter using its weight $w_i^{(c)}$:

$$\beta^{(c)}_{crit,i} = \frac{w_i^{(c)} * CG}{\max(NORU_i - sp_i, sp_i - NORL_i)} \tag{5}$$

In order to estimate the individual coefficient for the $i^{th}$ parameters, from a risk based approach, we divide by the longest distance from the set-point ($sp_i$) to the nearest NOR border: where $NORU_i$ is the upper boundary of the NOR and $NORL_i$ is the lower boundary of the NOR of the parameter $i$. Note that this works for a symmetric as well as asymmetric NOR.

3. Using the design matrix $X$, obtained for a specific experimental design, we can simulate possible $\tilde{y}$ values at the screening range using:

$$\tilde{y}^{(c)} = X\beta^{(c)}_{crit} \tag{6}$$

4. From that, the total sum of squares can be estimated:

$$\widehat{SS_{tot}}^{(c)} = \sum_i^n (\widehat{y_i \tilde{y}}^{(c)} - mean(\tilde{y}^{(c)}))^2 \tag{7}$$

Together with the sum of squares of the residuals the expected coefficient of variance can be calculated:

$$\tilde{R}^{2(c)} = 1 - \frac{\widetilde{SS_{res}}}{\widetilde{SS_{tot}}^{(c)}} \quad (8)$$

5. Using Choen's effect size ($f$), the non-centrality parameter λ and the critical F value ($F_{crit}$) the a-priori power for the combination c of effects that no parameter has been overlooked can be calculated [7]:

$$f^{2(c)} = \frac{\tilde{R}^{2(c)}}{1 - \tilde{R}^{2(c)}} \quad (9)$$

$$\lambda^{(c)} = f^{2(c)} * \nu \quad (10)$$

*1*

6. Confidence intervals for the a priori power for the combination c were calculated according to:

$$\lambda_{upp}^{(c)} = \lambda^{(c)} * c_{crit}(1 - \alpha|\nu)/\nu \quad (11)$$

$$\lambda_{low}^{(c)} = \lambda^{(c)} * c_{crit}(\alpha|\nu)/\nu \quad (12)$$

Where $c_{crit}(\alpha|\nu)$ is the $100 * \alpha$ percentile from a χ² distribution with ν degrees of freedom.

$$F_{crit} = F_{inv}(1 - \alpha|u, \nu) \quad (13)$$

7.

$$power_{apriori}^{(c)} = 1 - F_{nc}(F_{crit}|, u, \nu, \lambda^{(c)}) \quad (14)$$

Where $F_{nc}$ is the non-central F distribution with $u = p$(number of DoE parameters) and $\nu = n - u - 1$, where n is the number of observations in the DoE.

8. The mean power over all combination of effects was estimated as the arithmetic mean of all $power_{apriori}^{(c)}$:

$$power_{apriori} = \frac{\sum_{c=1}^{C} power_{apriori}^{(c)}}{C} \quad (15)$$

### 2.4. Evaluation of DoEs

Multiple linear models were used to identify the relationship of the studied PPs (DoE factors, X) on the response variable (y), representing a CQA or KPI of the process, up to a residual error (ε):

$$y = \beta_0 + X\beta + \varepsilon \quad (16)$$

Where X is a (n x p) dimensional design matrix for n DoE runs and p DoE factors which are studied, $\beta_0$ is the intercept, $\beta$ are the true effects of the DoE factors, and $\varepsilon$ is the residual, un-modelled error vector. The un-modelled error vector describes the analytical and process variance as well as non-linear effects which cannot be accounted for in the model structure. Identification of significant parameters was done using stepwise regression within multiple linear regression (MLR) tool of inCyght software (inCyght version 2017.03, Exputec GmbH). Parameters showing a partial p-value below 0.05 were allowed to enter the model. Those which have shown a p-value larger than 0.1 have been excluded from the model. Starting with the most significant parameter, this including/excluding procedure was applied iteratively and was repeated till the model structure did not change any more and the optimal model was achieved by this approach, identified significant parameters and their respective p-value are shown in Table 1 and Table 2 for CC 1 and PR, respectively. The normalized raw data are given in the supporting information Table S1 and Table S2 (section 9.3).

**Table 1.** p-values of significant process parameters that were used in the statistical models for each CQA of CC 1. Normal operating ranges and thresholds are given for each process parameter or critical quality attribute, respectively. Non-significant parameters are indicated with "-".Also the ratio of standard

deviation of raw residuals of the model by the standard deviation at set-point ($\sigma_{residues}/\sigma_{SP}$) is given for each CQA.

| | | End pooling [CV] | Elution strength [mM] | Wash Strength [mM] | Column loading density [g/L] | pH [-] | $\frac{s_{residues}}{s_{SDM}}$ |
|---|---|---|---|---|---|---|---|
| CQA | NOR[1] | -1.1 - 0 | -1.1 – 0.65 | -1.1 – 1.1 | -0.51 – 1.1 | -0.55 – 0.55 | |
| | Threshold | | | | | | |
| Process impurity 2 clearance | 0.85 | - | - | 0.059 | 0.099 | - | 7.79 |
| Product impurity 1 clearance | 1.08 | 0.028 | - | 0.098 | 0.089 | 0.027 | 18.12 |
| Product impurity 2 clearance | 0.1 | - | - | - | - | - | 256.06 |

[1] NOR was normalized by the screening range

**Table 2.** p-values of significant process parameters that were used in the statistical models for each CQA of PR. Normal operating ranges or thresholds are given for each process parameter or critical quality attribute. Non-significant parameters are indicated with "-". Also the ratio of standard deviation of raw residuals of the model by the standard deviation at set-point ($\frac{\hat{\sigma}_{residues}}{\hat{\sigma}_{SP}}$) is given for each CQA.

| | | Temperature [°C] | Time [hours] | Mixing [Yes/No] | pH [-] | $\frac{\hat{\sigma}_{residues}}{\hat{\sigma}_{SP}}$ |
|---|---|---|---|---|---|---|
| **CQA** | **NOR[1]** | -1.71 – 0.41 | 0.33 – 0.41 | -0.95 – 0.95 | -0.61 - 0.61 | |
| | **Threshold** | | | | | |
| **Process impurity 1 concentration specific** | $9*10^5$ | $9*10^{-5}$* | - | - | 0.07 | 64.89 |
| **Process impurity 2 concentration specific (prior filtration)** | $9*10^4$ | - | - | - | - | 2.68 |
| **Process impurity 2 concentration specific (post filtration)** | 784.7 | - | - | - | 0.021 | 0.55 |

[1] NOR was normalized by the screening range

* A quadratic effect was modelled for temperature and the shown p-value corresponds to the quadratic effect

## 3. Results & Discussion

Experiments performed in biotechnological studies might contain data that violate the statistical assumptions of parametric tests (i.e. normality, homogeneity of variances and independence of errors). Moreover, with a limited number of experiments and a large number of unknown parameters, such assumptions are hard to assess. Consequently, nonparametric approaches bear potential and we want

to present a novel permutation test to assess power of individual DoE factors in a multivariate regression model.

### 3.1. Permutation test for retrospective power analysis

The following permutation approach is adapted from a permutation test aiming to investigate power retrospectively [14]. Here we adapted this approach to study the significance of the alternative hypothesis that critical effects are present. Following steps are performed:

1. Using variable selection procedures, we select a significant regression model (all included effects are not 0 to a certain significance level):

$$y = \beta_0 + \beta_s * X_s + R_{y|X_s} \tag{17}$$

   where $X_s$ denotes the s significant parameters selected from a variable selection procedure (e.g. stepwise variable selection) and $R_{y|X_s}$ are the residuals of the obtained model. Find a list of those significantly selected parameters for the case studies of this work in Table 1 and Table 2.

2. We define a critical gap (CG) we must not surpass as the difference of the threshold and the worst case model prediction within the NOR ($x_{worst\ case_{NOR}}$), which is the parameter setting where the model prediction ($\hat{y}(x)$) is closest to the $threshsold_{USL}$:

$$CG = thesold_{USL} - \hat{y}(x_{worst\ case_{NOR}}) \tag{18}$$

3. Similar to the approach discussed in section 0 for the a priori power analysis, for non-significant parameters a variety of combinations (in total C) of effects for those parameters exist that lead to surpassing a critical threshold. In order to estimate the mean likelihood not to overlook a specific parameter, we vary the relative impact on the threshold of each parameter gradually between 0 and 1 in 100 steps. The fraction of the CG which is attributed to the non-significant parameter $i$ is expressed as the weight $w_i^{(c)}$ for the combination c. Eq. 5 can be used to calculate the critical effect of the parameter $i$.

4. The residuals $R_{y|X_s}$ are permuted randomly, producing $R^*{}_{y|X_s}$

5. New response values are calculated from the permuted residuals assuming that the critical effect is present under the alternative hypothesis ($H_A$):

$$y^* = \beta_0 + \beta_s * X_s + \beta^{(c)}{}_{crit} * Z + R^*{}_{y|X_s} \tag{19}$$

   where $\beta^{(c)}{}_{crit}$ is a vector of regression coefficients for the non-significant parameters and $Z$ is the design matrix for all non-significant parameters.

6. Make a model for $y^*$ based on X and Z and record significance of $\hat{\beta}_{crit}$ at a certain significance level (here α=0.05)

7. Repeat steps 4, 5 and 6 a large number of times (here 1000) and count number of significant outcome for each $\hat{\beta}_{crit,i}$ at a certain significance level (here α=0.05). The fraction of significant outcomes of all iteration cycles equals the retrospective power of parameter $i$.

### 3.2. Comparison of a priori and retrospective power

If we apply the proposed retrospective power analysis permutation test of section 0 to experimental data recorded from two unit operations (CC1 and PR) we obtain power values for each PP/CQA combination from Table 1 and Table 2, respectively.

Figure 1A shows a comparison of the retrospective and a priori power analysis for the CC1 unit operation. For all three studied CQAs at this stage ('process impurity 2 clearance', 'product impurity 2 clearance' and 'product impurity 1 clearance') we obtain a priori estimates of 1 (rightmost bar group in Figure 1A). This indicates an ideal case to start with experiments since there is no chance to overlook a critical effect. Retrospectively power analysis revealed that for all investigated PPs power values are well below the common statistical practice cut-off value of 0.8. This can be explained by the fact that the residual variance in the model is much higher than the initial estimate at set point, expressed by ratios of $\frac{\hat{\sigma}_{residues}}{\hat{\sigma}_{SP}}$ well above 1, as shown in Table 1. In general, multiple reasons for this discrepancy between the initial guess of expected variance and the actual residual variance in the model might exist. It could

69

be a non-representative selection of set-point runs (e.g. runs conducted with different operators), unexpected increase of variance during experiments (e.g. it is more difficult to control experiments at unusual parameter settings) or even non-linear dependency which cannot be captured by the linear model structure. Although statistically good practice, our experience shows that such non-linear dependencies might not be obvious from analysis of residuals (e.g. investigation of plots of residual vs. DoE factors). In a DoE approach, each experiment is unique in its settings if we do not use replicates and thereby no redundancy is available to hinder the model from being leveraged by non-linear responses.

For the precipitation step (PR) a priori power analysis again suggested a power of 1 (Figure 4B). Retrospectively assessed power values match the results obtained from a priori analysis, indicating that the performed DoE had sufficient power to assess critical effects of process parameters on quality attributes. This is reasonable since ratios of $\frac{\hat{\sigma}_{residues}}{\hat{\sigma}_{SP}}$ are closer to 1 for this unit operation compared to CC 1, as shown in Table 1.
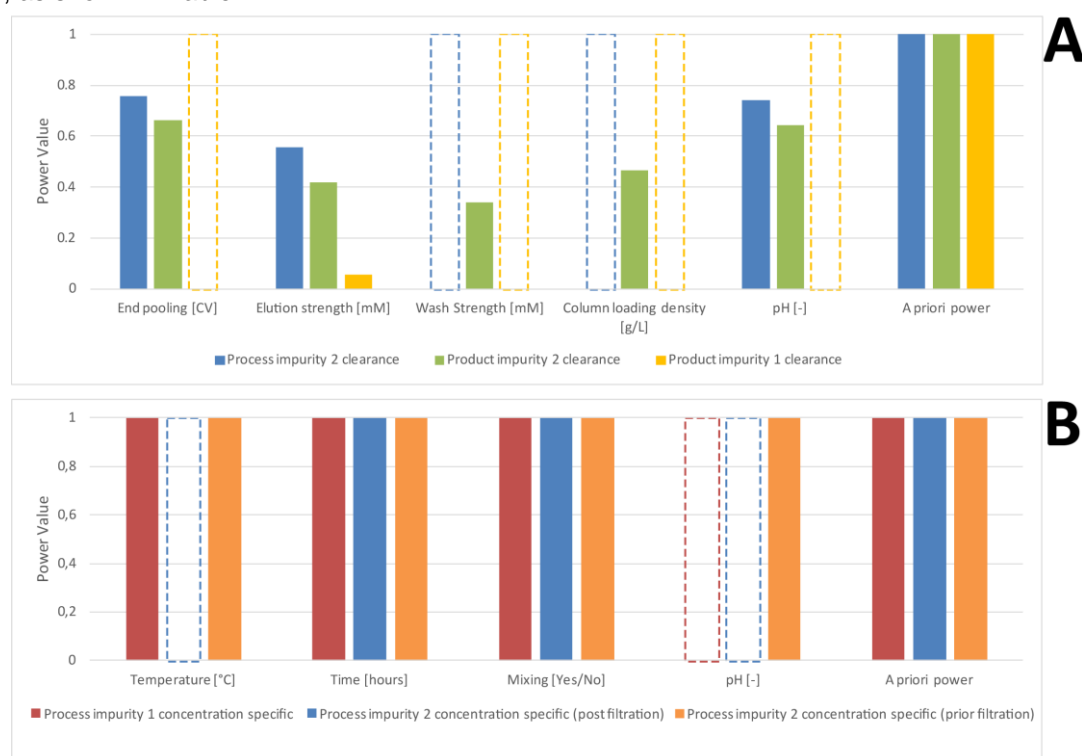


**Figure 1.** Power values for CC 1 (A) and PR (B) for each PP and CQA. Where significant process parameters were detected for a quality attribute, bars are marked grey. (A) Though a priori power analysis suggested a power of 100% for each investigated CQA for chromatography step 1, retrospective power analysis revealed that the power to detect a critical effect did not surpass 80 % for any of the investigated process parameters. Strategies to tackle these low-power-situations are given in Figure 4. (B) For the precipitation step, a priori power analysis suggested a power of 100% for each investigated CQA as well. Retrospective power confirmed these findings that is a 100 % chance that we did not overlook a critical effect of the investigated process parameters on quality attributes.

### 3.3. How to deal with low-powered parameters?

The most common approaches to tackle insufficient power values in screening designs are by increasing the sample size, reduction of measurement variance (either analytical or process), increasing the screening range if technically possible, or accepting the lack of power, however stating the parameter as key or critical. The latter strategy will have impact on the extended monitoring of such parameters during a subsequent process performance qualification (PPQ) campaign and routine manufacturing. As seen in section 0, a priori power analysis suggested high power values for all investigated unit operations, however, drastically overestimated the power for CC1. In specific cases, retrospectively increasing the sample size or the screening range might not be possible due to shortage

of starting material or technical limitations. A measurement method with less variance might not be at hand to re-measure backup samples. Another approach made possible by the presented method for power analysis is to narrow the NOR of some process parameters. If the threshold stays the same and the NOR is symmetrically located around the set point, for smaller NORs larger effect sizes are necessary to surpass the critical threshold as shown in Eq. 5 (i.e. steeper slopes). As a demonstrating scenario, we have chosen the relatively low power for Product impurity 2 clearance on CC 1 (see Figure 1A). For this response no significant parameter could be found. Figure 2 shows how a reduction of the NOR of the process parameter "wash strength" impacts the power of all studied PPs of this unit operation. Upon reducing the initially defined NOR by 50 % of its width the power value for "wash strength" increases from 0.34 to 0.68. As seen in Figure 2, the power values of the residual process parameters' effects on the same quality attribute remained unaffected neglecting the residual variation caused by the Monte Carlo approach in permutation.
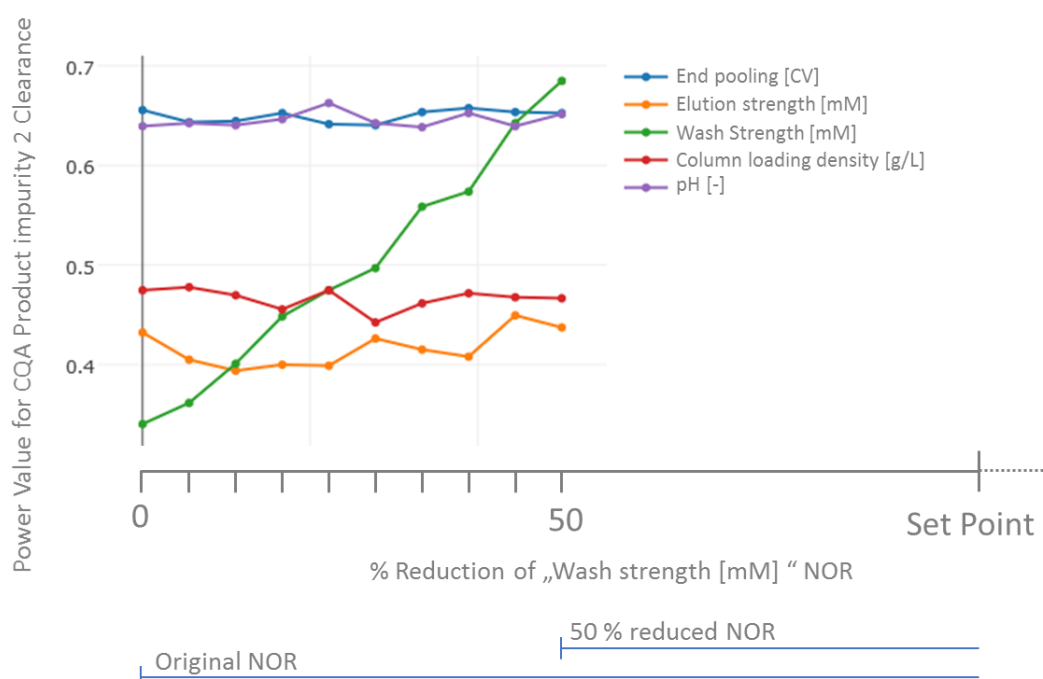


**Figure 2.** Retrospective power values for product impurity 2 clearance for unit operation CC1 as a function of tightened NOR of process parameter "wash strength". At the initially defined NOR the power value is 0.34. Upon reducing the NOR symmetrically by 50 %the power value for this process parameter increases to 0.68. The power values of the residual process parameters remain unaffected. The visible variation can be attributed to the variance in the permutation test.

This provides an opportunity to implement a tighter control strategy though adjusting the NOR as an approach to ensure no critical effects have been overseen. However, it may not be technically feasible or desirable for all process parameters to implement a tighter control strategy narrower ranges, especially for a parameter that has not been confirmed to significantly impact a CQA. Since a process parameter is studied in respect to multiple CQAs, we want to note that a tightening of a NOR of a process parameter that impacts significantly onto one specific CQA will also increase the power not to overlook this parameter regarding all other CQAs which have been studied in the same experiment. In contrast to changing the NOR of a non-significant parameter onto a CQA as shown for the combination 'wash strength' onto 'product impurity 2 clearance' in Figure 2, we investigated how a change of a significant parameter impacts on power levels. This was exemplarily done for decrease in NOR of 'wash strength' and we recorded power values for 'process impurity 2 clearance' of all non-significant parameters (here End pooling, elution strength and pH). We can see that due to the reduction of the NOR of a significant parameter the power values of all non-significant parameters increase too. In detail, a 50% reduction of

the NOR of the significantly impacting parameter "wash strength" increases the power of all non-significant parameters by approximately 10%. This can be explained by the fact that the worst case model prediction within the reduced NOR leads to a larger CG as defined in Eq. 18. Thereby, also the critical effects will be larger (Eq.6) and consequently the chances to overlook larger critical effects will be reduced. In this way an improved control strategy for a known significant parameter would improve the confidence that all residual non-significant parameter were not overlooked. This is potently a more desirable approach as improved control of known significant parameters is typically required and advantageous, if feasible.
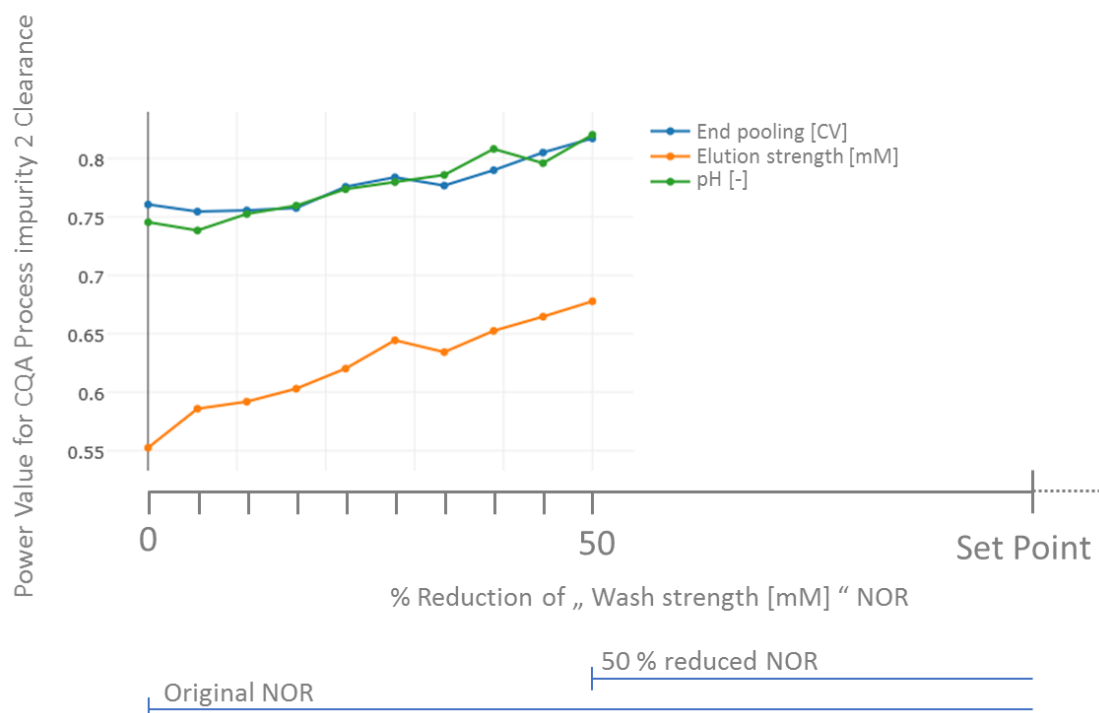


**Figure 3.** Retrospective power values for 'process impurity 2 clearance' for unit operation CC 1 as a function of tightened NOR of process parameter "wash strength". Since wash strength and column loading density are significant parameters in this model, the power was not assessed for those two parameters. Upon reducing the NOR symmetrically by 50 % of the significant parameter "wash strength" power values of all other parameters increase since the critical gap is increased, too, due to a reduction of the worst case model prediction in the NOR (Eq. 18).

### 3.4. Workflow for criticality assessment

In order to summarize the knowledge obtained from the application of the proposed posterior power analysis on two unit operations, we present a workflow that should aid process engineers in assessment of critical parameters (Figure 4). After selection of design and appropriate experiment number, a priori power analysis identifies if it is likely that a critical effect will not be overlooked. Sufficient power levels are normally assumed at 0.8 to 0.9. In cases where sufficient power cannot be assumed, the number of experiments, type of design or screening range must be increased. Both add to the expected signal to noise ratio. When increasing the screening range care must be taken not to obtain failure in experiments due to technical limitations or likely happening interaction effects (edge of failure). In order to reduce the risk of edge of failure experiments it is beneficial to conduct expected worst case setting of the process parameters first and potentially revise screening range afterwards.

In case sufficient power can be assumed, experiments can be conducted and regression modelling can be performed together with selection of significant DoE factors/parameters. After the "optimal" model was selected with its significant factors, retrospective power analysis as shown in section 0 will estimate chances that the residual non-significant factors might contribute to effects that surpass a pre-specified critical threshold. In case all non-significant parameters show power values well above 0.8 to 0.9, all of them can be stated as non-critical since the residual chance that they have been overlooked is only 20

to 10%, respectively. Otherwise, for those parameters that show insufficient power, analytical and/or reproducibility variance might be lowered by re-measurement of the samples or re-conducting of experiments, respectively. Another option is to narrow the NOR of potentially overlooked parameters which show large variability. This decreases their respective critical effect according to Eq. 5. After one of those three countermeasures has been taken retrospective power analysis can be repeated to ensure sufficient power values are reached and all parameters can be stated as non-critical. If none of the above three options is technically feasible or desirable, potentially overlooked parameters should be stated as critical and monitored during process performance qualification (PPQ) runs or routine manufacturing.
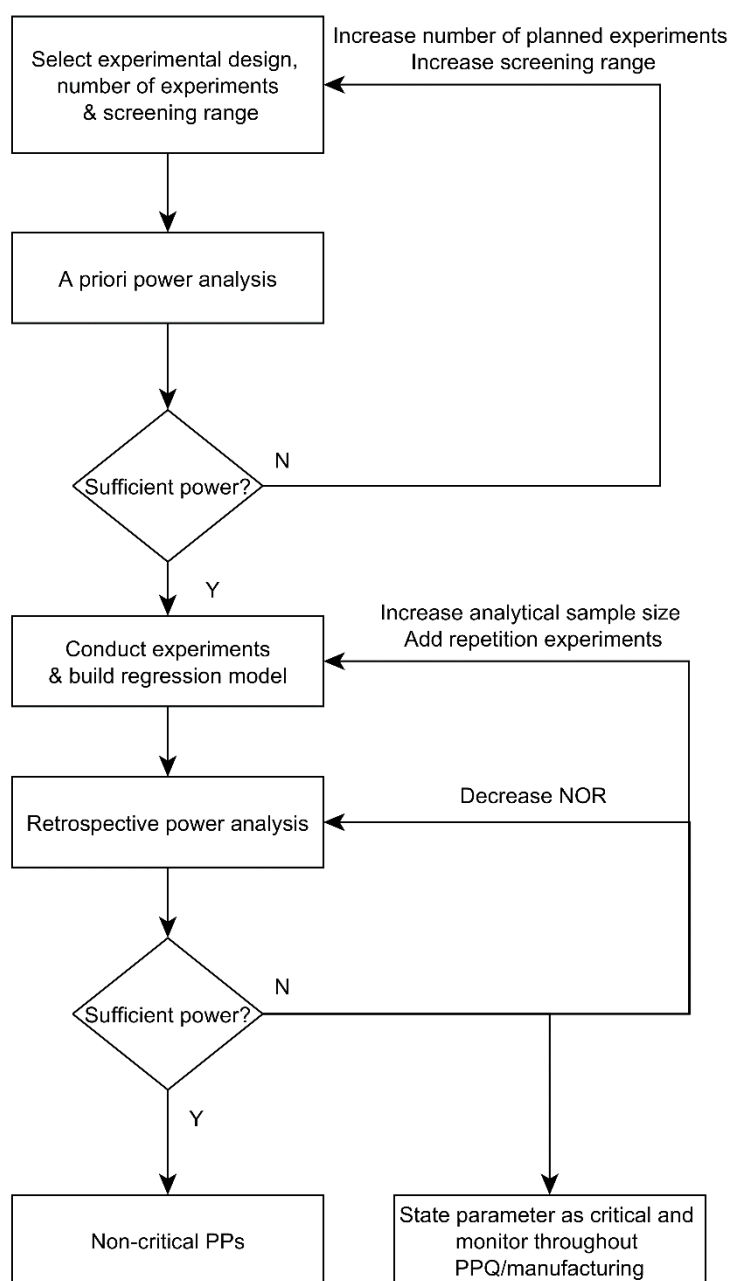


**Figure 4.** Workflow for criticality assessment of process parameters during process validation stage 1.

## 4. Conclusion

The goal of the contribution was to demonstrate the capability of a multivariate retrospective power analysis methodology to identify critical process parameters during pharmaceutical process validation stage 1.

We have shown in a case study that parameters that are non-significant in models, which have initially thought to be sufficiently powerful to identify critical effects, might still show effects that surpass a critical threshold due to increased analytical, process, or reproducibility variance. This leads to situations where impact of those parameters on final drug product quality cannot be excluded. This was shown using a biopharmaceutical case study conducted at a world leading CMO. However, common practice is to state such parameters as non-critical and thereby overlook their potential harmful impact. Therefore, two missing parts have been introduced in this contribution: (i) a novel permutation methodology for multiple linear regression that estimates retrospective power (i.e. the chance of non-significant parameters to mutually combine to a critical effect) and (ii) a workflow for criticality assessment that shows strategies how to mitigate the risk of low-powered parameters. Besides the well-known fact that an increase in experiments increases power, it could be shown that a reduction of the NOR of significant parameters increases power of all non-significant parameters via a reduction of the worst case model predictions. While a reduction of the NOR of a specific non-significant parameter increases power solely for this parameter. Additionally, if implementation of tighter NOR controls is practically infeasible, this methodology can at a minimum appropriate assess the process risk and increase awareness to the limitations of the initial classification, potentially suggesting an improved control strategy is required.

Using both tools, it will be possible for process engineers during the design stage of a process validation (stage 1) to:

- reduce the chance of overlooking potentially CPPs
- develop a control strategy for potentially overlooked CPPs in order to increase process robustness
- lower OOS events and finally contribute to increased patient safety.

## References

1. Ahir, K. B.; Singh, K. D.; Yadav, S. P.; Patel, H. S.; Poyahari, C. B. Overview of Validation and Basic Concepts of Process Validation. *Sch. Acad. J. Pharm.* **2014**, *3*, 178–90.
2. FDA Guidance for Industry. **2011**.
3. Katz, P.; Campbell, C. FDA 2011 process validation guidance: process validation revisited. *J. GXP Compliance* **2012**, *16*, 18.
4. Mollah, A. H. Application of failure mode and effect analysis (FMEA) for process risk assessment. *BioProcess Int.* **2005**, *3*, 12–20.
5. Guideline, I. H. T. Pharmaceutical Development Q8 (R2). *Curr. Step* **2009**, *4*.
6. Peres-Neto, P. R.; Olden, J. D. Assessing the robustness of randomization tests: examples from behavioural studies. *Anim. Behav.* **2001**, *61*, 79–86, doi:10.1006/anbe.2000.1576.
7. Cohen, J. *Statistical power analysis for the behavioral sciences*; Rev. ed.; Academic Press: New York, 1977; ISBN 978-0-12-179060-8.
8. Nickerson, R. S. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol. Methods* **2000**, *5*, 241.
9. Thomas, L. Retrospective Power Analysis. *Conserv. Biol.* **1997**, *11*, 276–280, doi:10.1046/j.1523-1739.1997.96102.x.
10. Thomas, L.; Krebs, C. J. A review of statistical power analysis software. *Bull. Ecol. Soc. Am.* **1997**, *78*, 126–138.
11. Jones, B.; Nachtsheim, C. J. A class of three-level designs for definitive screening in the presence of second-order effects. *J. Qual. Technol.* **2011**, *43*, 1.

12. Tai, M.; Ly, A.; Leung, I.; Nayar, G. Efficient high-throughput biological process characterization: Definitive screening design with the Ambr250 bioreactor system. *Biotechnol. Prog.* **2015**, *31*, 1388–1395, doi:10.1002/btpr.2142.
13. Kraber Shari; Whitcomb Pat; Anderson Mark Handbook for Experimenters.
14. Freedman, D.; Lane, D. A Nonstochastic Interpretation of Reported Significance Levels. *J. Bus. Econ. Stat.* **1983**, *1*, 292–298, doi:10.2307/1391660.

# Integrated Process Modeling – A process validation life cycle companion

**Thomas Zahel [1], Stefan Hauer [1], Eric M. Mueller [2], Pat Murphy [2], Sandra Abad [3], Elena Vasilieva [3], Daniel Maurer [3], Cécile Brocard [3], Daniela Reinisch [3], Patrick Sagmeister[1] and Christoph Herwig [1,*]**

[1]  Exputec GmbH, Mariahilferstraße 147, Vienna, Austria; thomas.zahel@exputec.com
[2]  Versartis Inc., 4200 Bohannon Drive, Suite 250 Menlo Park, California, USA; emueller@versartis.com
[3]  Boehringer Ingelheim RCV GmbH & Co KG, Doktor-Boehringer-Gasse 5-11, Vienna, Austria; sandra.abad@boehringer-ingelheim.com
*  Correspondence: christoph.herwig@exputec.com; Tel.: +43-1-9972849

**Abstract:** During the regulatory requested process validation of pharmaceutical manufacturing processes, companies aim to identify, control and continuously monitor process variation and its impact on critical quality attributes (CQAs) of the final product. It is difficult to directly connect the impact of single process parameters (PPs) to final product CQAs, especially in biopharmaceutical process development and production, where multiple unit operations are stacked together and interact with each other. Therefore, we want to present the application of Monte Carlo (MC) simulation using an integrated process model (IPM) that enables estimation of process capability even in early stages of process validation. Once the IPM is established, its capability in risk and criticality assessment is furthermore demonstrated. IPMs can be used to enable holistic production control strategies that take interactions of process parameters of multiple unit operations into account. Moreover, IPMs can be trained with development data, refined with qualification runs, and maintained with routine manufacturing data which underlines the lifecycle concept. These applications will be shown by means of a process characterization study recently conducted at a world leading contract manufacturing organization (CMO). The new IPM methodology therefore allows to anticipate out of specification (OOS) events, identify critical process parameters, and take risk based decisions on counteractions that increase process robustness and decrease the likelihood of OOS events.

**Keywords:** process validation; process characterization study; holistic process model; predict out of specification events; Monte Carlo simulation; biopharmaceutical manufacturing

## 1. Introduction

The main goal of pharmaceutical manufacturing is to constantly deliver high product quality, which is reflected in regulatory guidelines [1–3]. Process validation is a major initiative to demonstrate the capability of meeting this goal and is separated in three stages (stage 1 to 3). Stage 1 aims at establishing a process design in which process variation in critical quality attributes (CQAs) is understood and connected to critical process parameters. This is usually done within a process characterization study using design of experiment (DoE) strategies. Resulting critical process parameters that have an effect on product quality require sufficient control strategies. Stage 2 consists of process performance qualification (PPQ) runs to confirm the design of the process and ensure it can consistently deliver high product quality. Stage 3, continued process verification (CPV), is an ongoing evaluation and monitoring to confirm the process remains in a state of control or to identify if new interdependencies between process parameters (PPs) and CQAs arise. Those three stages can be seen interlinked to each other as a lifecycle, where potential changes and associated risk in PPQ or routine manufacturing must be iteratively evaluated together with knowledge gained from initial process design [4]. Insufficient risk estimation of the entire process at stage 1 of process design (e.g. in terms of

estimation of out of specification events) can lead to inconsistent or unpredicted performance at later stages.

Risk evaluation of individual unit operation of a pharmaceutical processes is commonly conducted by following steps in accordance with ICH Q8 guideline [2]:

- Risk assessment using knowledge of process experts, which leads to a candidate set of potential critical PPs for each unit operation.
- Experimental investigation of the impact of potentially critical PPs onto CQAs. This is usually performed in DoE approaches and statistical regression modelling is used to describe the relationship between significantly impacting critical PPs and CQAs mathematically.
- Comparison of the output of statistical model predictions within normal operating ranges or a design space to pre-defined acceptance limits for each unit operation.
- The risk of not meeting acceptance limits is mitigated by applying an appropriate control strategy, such as a reduction of the normal operating range.

One difficulty, especially in biopharmaceutical manufacturing where multiple unit operations are stacked together and critical PPs interact, is an appropriate evaluation of risk related to impurities. Risk analysis is impeded since propagation of impurities is rarely assessed holistically but rather evaluated on each unit operation separately [5]. Impurity propagation through multiple unit operations is difficult to study with reasonable representative experimental effort, especially at early stages of process design where only a limited number of manufacturing runs is available. However, simulations and modeling are necessary and useful to assess the chance of out of specification events. Having such a predictive tool in place to develop robust processes by incorporating knowledge acquired during process development and characterization experiments, unexplained variance in product quality possibly leading to recalls, complaints, and patient risk can be reduced. Therefore, it is desirable to formulate holistic process and production control strategies that prevent out of specification (OOS) events which could have already been anticipated within the design phase [6]. However, to the best of our knowledge, it has not been shown so far how a holistic risk evaluation spanning over multiple unit operation can be performed at process validation stage 1 and used to demonstrate overall process capability.

MC simulation is a tool to incorporate random variability into the modelled system and connect single modelling-units together. A random sampling distribution for the model parameters (inputs) needs be defined a priori, which does not need to be necessarily normally distributed. Within each cycle of the MC simulation a different random set of inputs is drawn leading to discrete model results (outputs). Since a large number of MC cycles are performed it is possible to aggregate the discrete model outputs to a predictive distribution of those outputs. Using this distributional information it is possible to calculate probabilities of events (e.g. OOS). MC simulations have shown great potential in pharmaceutical industry for drug discovery and simulation of clinical trials [7] and is also routinely utilized for error propagation [8]. However, it has, to our knowledge, not been applied to impurity propagation of a batch-wise pharmaceutical processes.

Herein, we describe the development of an integrated process model (IPM) that is capable of capturing development and design data from multiple unit operations and is able to predict the risk of OOS probabilities through Monte-Carlo simulation even at the early design stage of process validation. Moreover, we identify how variance and changes in set point of process parameters impacts drug substance quality. The model can be enriched at later stages also with data from PPQ, routine manufacturing, or additional development. Thereby a continuous process data management is enabled and risk based decision making during change and deviation management in continuous manufacturing can be based on the full spectrum of development, design, and manufacturing data.

At this stage, the following derived acceptance criteria for the IPM can be formulated:

- Prove process robustness of an existing design space: Prove that under normal manufacturing conditions it is unlikely to miss drug substance specification for defined CQAs
- Test process robustness under accelerated variance of process parameters and increased impurity burden
- Establish a platform that leverages process knowledge from PV stage 1 for further usage within PPQ and CPV (Stage 2 and 3 of process validation)

With this contribution we present the development of an IPM, validate the IPM using large scale manufacturing data, and demonstrate the capability of the IPM in estimating OOS probabilities under

normal and accelerated conditions. This case study was recently conducted at a leading biopharmaceutical CMO in contract development of a novel long acting human growth hormone product.

## 2. Materials and Methods

Here we want to summarize the required inputs for the IPM as well as their assumptions that must be met in order to ensure reliable prediction of the IPM (for details see referred sections):

- Description of the process, order of unit operations and variance of PPs under normal operating conditions (see section 2.1). It is assumed that estimation of variance of PPs is representative for routine manufacturing.
- Optional: If initial unit operation of the process are not modelled by the IPM the starting distribution of each CQA needs to be estimated at the starting unit operation of the IPM. It is assumed that the estimation of starting distribution is representative for the real CQA distribution under routine manufacturing (see section 2.2).
- Statistical regression models that describe significant relationships between PPs and CQAs for each unit operation (see section 2.3.1). It is assumed that scientifically sound analytical methods (high accuracy, precision, robustness, selectivity, etc.) have been used to record the data that led to formation of those regression models. Moreover, it is assumed that no critical effect has been overlooked, which can be tested using power analysis approaches [9]. This ensures that residual variance in the regression models can be attributed to normal analytical- and process variance.
- Optional: Statistical spiking models of each unit operation describing the dependency between varied impurity load and specific impurity clearance (see section 2.3.2). Identical assumptions as for the regression models must be met.

### 2.1. Description of biopharmaceutical manufacturing process

This industrial biopharmaceutical process produces a pharmaceutically active recombinant protein and is divided into 7 unit operations. After fermentation using *Escherichia coli* as host cells and recombinant expression of the product, a cell lysis step is performed prior to a precipitation step and clarification. After these primary recovery steps, three preparative chromatographic columns are performed to clear the product from impurities. A final ultrafiltration/diafiltration is performed to adjust the product concentration in drug substance. Two process related impurities as well as 2 product related impurities were defined as the major CQAs and herein modelled within the IPM. Since the analytical quantification of those CQAs was only possible in the load of the first chromatographic step, this step was set as input to the IPM. A summary of the relevant unit operations for modelling, their varied PPs within DoEs, the relative standard deviation of those PPs between large scale (LS) runs and the monitored CQAs is given in Table 1.

**Table 1.** Available data sets, process parameters and monitored CQAs for each unit operation included in the IPM. CC is abbreviation for chromatography column, PCI stands for process related impurities and PRI product related impurities.

| UO | Available Data sets | PPs varied in DoEs | Rel. std. of PPs between LS [%][1] | Std / NOR [%][2] | Monitored CQAs |
|---|---|---|---|---|---|
| CC 1 | • 9 manufacturing runs<br>• 13 DoE runs with definitive screening design | pH [-] | 1.61 | 46 | PCI 1, PCI 2, PRI 1, PRI 2 |
| | | Column loading density [g/L] | 12.05 | 50 | |
| | | Wash Strength [mM] | 5.00 | 62 | |

| | | Parameter | 1 | 2 |
|---|---|---|---|---|
| | | Elution strength [mM] | 5.00 | 44 |
| | | End pooling [CV] | 1.36 | 40 |
| CC 2 | • 9 manufacturing runs<br>• 11 DoE runs using full factorial design<br>• 1 spiking run with increased PRI 1 concentration in load<br>• 1 spiking run with increased PCI 1 concentration in load | pH [-] | 0.79 | 30 |
| | | Column loading density [g/L] | 4.84 | 20 |
| | | Gradient slope [% of Buffer] | 5.00 | - |
| CC 3 | • 9 manufacturing runs<br>• 9 DoE runs using definitive screening design | pH [-] | 0.92 | 35 |
| | | Column loading density [g/L] | 12.78 | 30 |
| | | Gradient slope [% of Buffer] | 5.00 | - |
| | | Wash Strength [mM] | 5.00 | 50 |

[1] Relative standard deviation to the set-point of the process parameters;

[2] Ratio of one standard deviation to the normal operating range.

### 2.2. Scope of IPM and Sampling Distribution of PPs

Due to limited amount of quantitative analytical data of the CQAs before chromatography column 1, the starting distribution of each CQA at the first chromatography step was assumed to be normally distributed with mean and standard deviation estimated from measured CQA distribution of LS runs. From this starting point the following unit operations chromatography column 1, 2, and 3 were modelled. The pool of chromatography column 3 was regarded as very similar to drug substance since no further clearance formation was expected at the ultrafiltration/diafiltration step.

For the MC workflow, we have to choose a realistic distribution of the large scale variation in process parameters in order to incorporate process related variability. Results of the MC simulation are dependent on the sampling strategy for the process parameters at each simulation. Often pseudo-random numbers are replaced by quasi-random numbers or Latin hypercube sampling [10,11] for better overview of possible outcomes. However, for realistic risk assessment, we want our sampling to be representative for the process, therefore classical pseudo-random numbers were used for sampling. Existing variance of process parameters has been estimated from current large scale runs as listed in Table 1. We assumed a multivariate normal distribution for all process parameters centered at their mean (target of operation) and variance from large scale runs without any covariance between the process parameters. This is a suitable simplification since process parameters are controlled independently from each other. In general, this is not a prerequisite for the IPM and might be adapted for other processes, where additional information of potential correlation between the process parameters exists.

### 2.3. Impurity Clearance Models

Since it was aim of the IPM to model the final distribution of each of the four major CQAs (i.e. the specific concentration of each impurity) and the product in the drug substance, their reduction from load of chromatography column 1 until drug substance needs to be described mathematically. In order to

estimate the specific CQA concentration after a unit operation (pool) using the specific load concentration of this CQA, specific clearances (SCs) were used:

$$Specific\ Clearance = SC = \frac{c_{CQA,load}}{c_{CQA,pool}}, \tag{1}$$

where $c_{CQA,load}$ and $c_{CQA,pool}$ is the specific CQA concentration, defined as the amount of impurity per amount of product, for load and pool, respectively.

For modelling the product a similar approach was chosen using step yields (SY) instead of SC:

$$StepYield = SY = \frac{p_{pool}}{p_{load}}, \tag{2}$$

where $p_{pool}$ and $p_{load}$ are the amounts of product in pool and load, respectively.

Two major impacting sources specific clearances have been considered here: (i) Impact of potential critical process parameters, which have been purposefully selected in a risk assessment and (ii) specific amount of impurity load per column volume. Those types of models are described in more detail in the following two sections 2.3.1. and 2.3.2., respectively. In case it was not possible to find any PPs that significantly impact on the clearance, the mean clearance from LS was taken as a constant model (see section 2.3.1 for details). We summarize all found models in Table 2.

**Table 2.** Summary of the presence of models that describe the relationship of a CQA specific clearance factor as a function of PPs (indicated by "DoE") or the impurity loading density of the respective CQA ("Spiking") and the respective p-value of the regression. In cases where no significant function of PPs on a CQA clearance could be found, mean large scale clearance was assumed indicated by "LS clearance" in the table. CC is abbreviation for chromatography column, PCI stands for process related impurities and PRI product related impurities.

| CQA/Unit Operation | CC 1 | CC 2 | CC 3 |
|---|---|---|---|
| PRI 1 | DoE (linear, p=0.09) | LS clearance + Spiking (p=0.00) | DoE (quadratic, p=0.01) |
| PRI 2 | DoE (linear, p=0.01) | LS clearance | LS clearance |
| PCI 1 | DoE (quadratic, p=0.00) | LS clearance + Spiking (p=0.04) | DoE (quadratic, p=0.00) |
| PCI 2 | LS clearance + Spiking (linear, p=0.00) | LS clearance | LS clearance+ Spiking (linear, p=0.00) |
| Yield | DoE (linear, p=0.00) | LS clearance | DoE (quadratic, p=0.00) |

2.3.1. Clearance and yield as a function of process parameters (DoE models)

As a general good practice in PV stage 1, after a purposeful selection of potential impacting process parameters, their impact on the SCs and the SY has been tested within DoEs. For reasons of simplicity we will only show the modelling approach for SC in the following two sections and not for step yields, since both approaches are identical when exchanging SC with step yield. Experimental designs were chosen (see Table 1 for number of DoE runs and design) and linear models were established according to the form:

$$SC = PP * \beta_{PP} + \beta_0 + \varepsilon, \tag{3}$$

where $SC$ is a (n × 1) vector of the measured specific clearances, $PP$ is a (n × p) matrix of the process parameter settings of each DoE run, $\beta_{PP}$ are the regression coefficients and $\beta_0$ is the intercept. The process of selecting a subset of significant process parameters was accomplished by means of stepwise regression using multiple linear regression (MLR) package in inCyght (inCyght 2017.03, Exputec GmbH, Vienna, Austria). In this stepwise procedure parameters showing a partial p value below 0.05 were allowed to enter the model starting with those parameters having the lowest partial p value. Partial p values of parameters can change as other parameters are included in a multivariate model. Therefore, after each time including a new parameter in the model, it is checked if p values of the existing parameters have increased and those parameters showing an p value larger than 0.1 will be excluded

80

from the model. This including/excluding procedure was applied iteratively to achieve the optimal model, starting with the most significant parameter and was repeated as long as the model structure did not change any more. Thereby, $\hat{\beta}_{PP}$ and $\hat{\beta}_0$ could be estimated. The herein obtained models and their respective statistical quality measures are summarized in Table S1 (section 9.4) of the supplementary materials.

A new prediction for SC ($\widehat{SC}$) for randomly selected set of process parameters of the $i^{th}$ MC simulation can be obtained by:

$$\widehat{SC}(PP^{(i)}) = PP^{(i)} * \hat{\beta}_{PP} + \hat{\beta}_0, \tag{4}$$

The prediction error of the mean model response was assumed to be normal distributed with: $N\left(\widehat{SC}(PP^{(i)}), \sigma^2{}_{\widehat{SC}(PP^{(i)})}\right)$. Where $\sigma_{\widehat{SC}(PP^{(i)})}$ was calculated using:

$$\sigma_{\widehat{SC}(PP^{(i)})} = s_{SC} * \sqrt{\frac{1}{n} + h_i}, \tag{5}$$

with the leverage of the new data point: $h_i = diag(PP^{(i)}(PP'PP)^{-1}PP^{(i)'})$, the residual standard error: $s_{SC} = \sqrt{\frac{\sum(SC_i - \widehat{SC_i})^2}{n-p-1}}$ if p are the number of parameters and n the number of observations. A random sample $rand(N\left(\widehat{SC}(PP^{(i)}), \sigma^2{}_{\widehat{SC}(PP^{(i)})}\right))$, using MATLAB (MATLAB, The MathWorks Inc. R2015b) function randn, was taken from this prediction error distribution for each Monte-Carlo simulation i and added to the mean prediction, obtaining the specific clearance impacted by PPs for each unit operation:

$$\widehat{SC}(PP^{(i)}) = rand(N\left(\widehat{SC}(PP^{(i)}), \sigma^2{}_{\widehat{SC}(PP^{(i)})}\right)), \tag{6}$$

For responses where no further spiking models have been available, the specific CQA concentration of the pool of the $u^{th}$ unit operation was calculated to:

$$c_{CQA,pool,u}{}^{(i)} = \frac{c_{CQA,load,u}{}^{(i)}}{\widehat{SC}(PP^{(i)})} = \frac{c_{CQA,pool,u-1}{}^{(i)}}{\widehat{SC}(PP^{(i)})}, \tag{7}$$

Note, that here the concatenation of the unit operations occurs since the specific CQA concentration of the pool of unit operation $u - 1$ is set equal to the load of unit operation $u$.

If no significant effects of any PP on an impurity clearance of a certain unit operation could be detected, a constant impurity clearance was assumed within the entire design space. This was modelled by the mean clearance of the LS runs and variance of the LS runs. In those cases, for each unit operation the specific clearance of the $i^{th}$ MC simulation reduces to:

$$\widehat{SC}(PP^{(i)}) = rand(N(\overline{SC}_{LS}, \sigma^2{}_{SC_{LS}})), \tag{8}$$

where $\overline{SC}_{LS}$ and $\sigma^2{}_{SC_{LS}}$ is the mean SC and the variance from LS runs, respectively.


2.3.2. Increased clearance due to varied spiking of impurities

During process development and design, increased impurity levels were spiked on chromatographic preparative columns in order to show elevated clearance capacity. In more detail, during those spiking studies, it was shown that the impurity clearance increases with increasing impurity loading density ($ILD = \frac{loaded\ impurity\ amount}{column\ volume}$) up to the tested level. Additionally, the same relationship of increased impurity clearance at increased impurity loading densities was found for large scale runs, where the impurity loading varies for each run due to variation in fermentation and previous purification unit operations. Since the ILDs were not included within DoE approaches as an independent DoE factor, we followed a two-step approach to incorporate altered clearance at varying ILD.

In the first step, linear regression on SC as a function of ILD was applied to identify significant correlations. Having such a regression model in place, for a specific ILD in the $i^{th}$ MC simulation an estimate for the SC could be obtained ($\widehat{SC}(ILD^{(i)})$):

$$\widehat{SC}(ILD^{(i)}) = rand\left(N\left(\widehat{SC}(ILD^{(i)}), \sigma^2{}_{\widehat{SC}(ILD^{(i)})}\right)\right), \tag{9}$$

where $\widehat{SC}(ILD^{(i)})$ is the mean predicted SC from the linear regression model at the specific $ILD^{(i)}$ and $\sigma^2_{\widehat{SC}(ILD^{(i)})}$ is the variance of the mean prediction, which can be obtained analogous to Eq. 5. An example of such a spiking model is shown in Figure 1, where an increased loading density of process related impurity 2 shows a significant (p=7×10-8) increase in specific clearance of process related impurity 2. Significant (p-value < 0.05 as well as R² (explained variance) - Q² (from leave one out cross validation) difference < 0.3) spiking models were selected for each response/unit operation and are summarized in Table 2 and Table S2 of the supplementary materials (section 9.4).
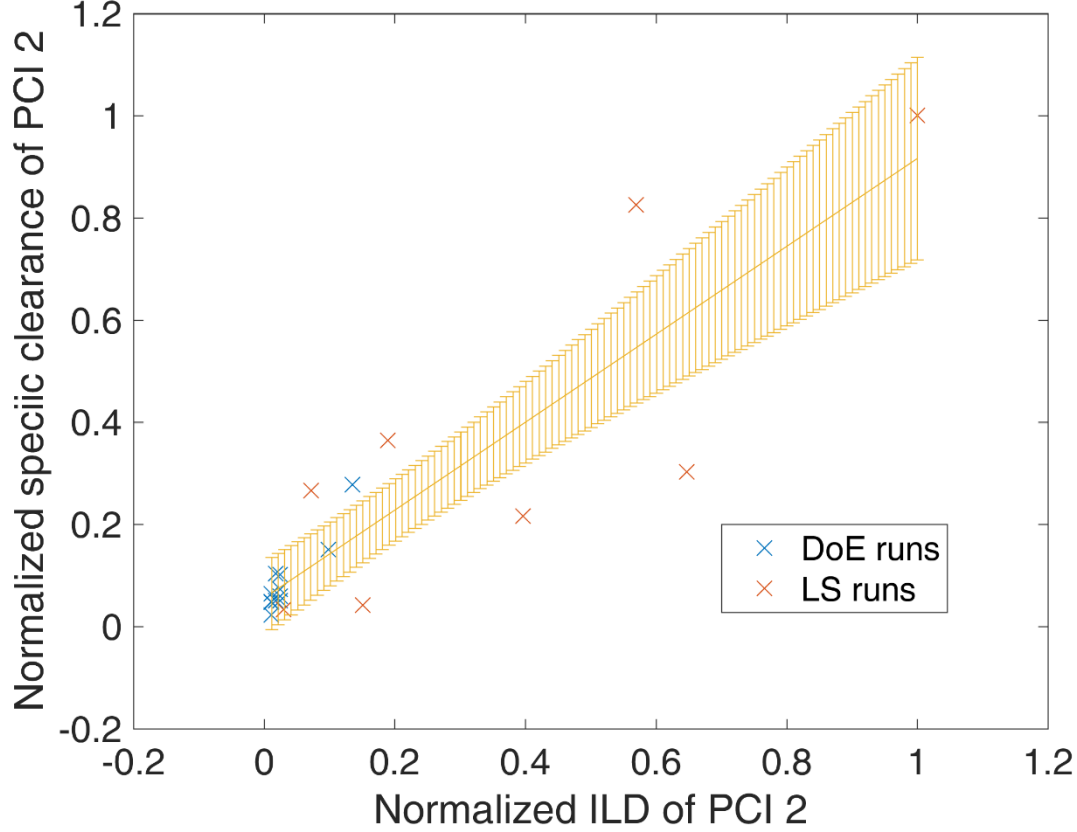


**Figure 1.** Exemplary plot for dependency of specific clearance (here of process related impurity 2) against impurity loading density of process related impurity 2 of DoE runs (blue) and large scale (LS) runs (red). Yellow error bars indicate the mean model prediction error. Normalization has been performed by division of the maximal value for each axes.

Hereafter in the second step, if significant spiking models were available, they were combined with the existing ones as a function of PPs as described in section 2.3.1. Therefore, for each unit operation, we added the expected clearance increase due to increased ILD to the specific clearance of the i-th Monte Carlo simulation impacted by PPs:

$$\widehat{SC}^{(i)} = \widehat{SC}(PP^{(i)}) * \frac{\widehat{SC}(ILD^{(i)})}{\widehat{SC}(\overline{ILD})}, \tag{10}$$

where $\widehat{SC}(\overline{ILD})$ is the SC under mean ILD from DoE runs. The ILD of the simulation i and the unit operation u can be calculated according to:

$$ILD_u^{(i)} = \frac{c_{load,u}^{(i)} * p_{load,u}^{(i)}}{CV} = \frac{c_{pool,u-1}^{(i)} * p_{pool,u-1}^{(i)}}{CV}, \tag{11}$$

where $c_{load,u}^{(i)}$ is the specific concentration of the CQA at the i-th simulation and the u-th unit operation and $p_{load,u}^{(i)}$ is the product amount modelled by step yield of simulation i and unit operation u, $CV$ is the column volume. Again, the load concentrations and amounts can be expressed by the respective pool concentrations of the previous unit operation (u-1).

82

Since the impurity loading density was not included within DoE approaches on column steps as an independent DoE factor, we assume that varied impurities do not show interactive effects with other DoE factors (mainly process parameters) within normal operating variance. In order to estimate the risk that the simulation performance is biased by the spiking models and the risk of the above stated assumptions, the IPM was simulated without applying any spiking model. Those results are shown in Figure S1-S4 of the supplementary materials (section 9.4), where we show that only for product related impurity 1, process related impurity 1 and process related impurity 2 the out of specification chance increases by 0.1, 0.7 and 4.2 %, respectively. Therefore, the above mentioned assumptions about spiking models and the connection to DoE models can be seen as a minor influence to the overall IPM prediction and valid simplification. Moreover, this can be regarded as a valid simplification since the assumed normal manufacturing variance which is used during IPM simulation of process parameters is well within the normal operating range (NOR, see standard deviation to NOR ratio in Table 1 is often below 30 %) and therefore around 99% the simulated batches are run within NOR. However, we want to note that one could even refine the IPM by including specific impurity concentrations in the load of chromatographic columns as an additional factor in DoE experiments to study that effect in combination with all other DoE factors.

## 3. Results

### 3.1. Monte Carlo Approach for integrated process modelling

The main idea behind the integrated process is to concatenate impurity clearance models of each unit operation together to predict the CQA distribution at each intermediate and at drug substance. To account for error propagation during this concatenation we performed a Monte-Carlo approach in four steps:

1. 1000 simulations were performed each having a different set of PPs ($PP^{(i)}$) for the three modelled unit operations (chromatography column 1/2/3) and different initial specific CQA concentrations ($c^{(i)}_{CQA,init}$) at the load of chromatography column 1, sampled from distributions which were estimated from LS runs. Also the variance in PPs was estimated from LS runs and is indicated by a schematic distribution on the x-axis in Figure 2A and Figure 2B. Additional increase in simulations did not increase model accuracy and 1000 simulations are a common standard for Monte Carlo simulations[7]. A more detailed description of this step and a list of used process parameters are provided in section 2.2.

2. For each unit operation, we modelled the specific clearance (SC) of each CQA as a function of the critical PPs and the ILD by multiple linear regression. Each model is associated with a prediction error, which is indicated by the blue shaded area around the found regression line Figure 2A and Figure 2B. The ILD can be derived from $c_{CQA,load}$ of each unit operation, which equals $c_{CQA,init}$ for the first modelled unit operation and $c_{CQA,pool,u-1}$ for all subsequent modelled unit operations (u).

3. Since $c_{CQA,pool,u}$ can be calculated from SC and $c_{CQA,load,u}$, on the whole, $c_{CQA,pool,u}$ can be seen as a function of $PP_u$ as well as $c_{CQA,init}$ or $c_{CQA,pool,u-1}$ as indicated in the formula of Figure 2A Figure 2B, respectively. Thereby the model outputs from multiple unit operations can be stacked together, which is indicated by black arrows in Figure 2A more thorough description of which models could be found on which CQA and unit operation is depicted in section 2.3.

4. Since we performed 1000 simulations, each having different settings in process parameters, we obtained a distribution for the specific CQA concentration in the pool and finally in drug substance, indicated on the y-axes of Figure 2A and Figure 2B and by the distribution in Figure 2C.
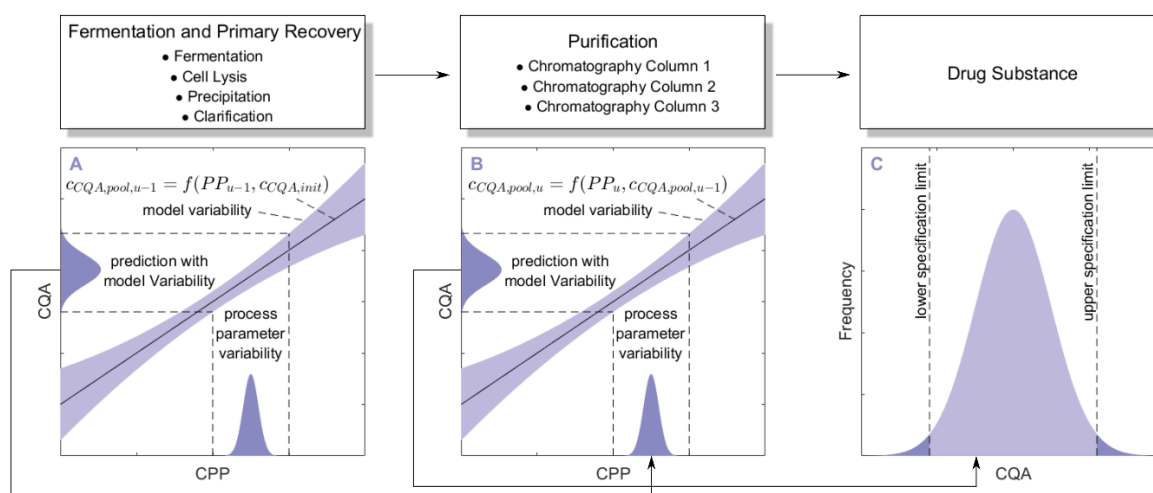
**F**igure 2. Schematic description of the integrated process model using a Monte-Carlo approach: 1000 simulations are performed, each having a different set of process parameters (indicated as distribution on the x-axes of A and B) and initial specific CQA concentration ($c_{CQA,init}$). Multiple linear regression models describe the relationship between the $c_{CQA}$ of the pool of unit operation u (B) and the PP of this unit operation as well as the pool concentration of the previous unit operation u-1 (A). Thereby models from multiple unit operations (A and B) are connected to predict the CQA distribution in the drug substance (C). Since 1000 simulations are performed, the CQA values form a distribution after each unit operation. The higher the model uncertainty, indicated by blue shaded area around the regression line, the wider the resulting CQA distribution. This ultimately propagates until drug substance, where the chance of out of specification events can be assessed.

### 3.2. Validation of the IPM using Observed CQA distribution in Drug Substance

For model validation, the distribution of the predicted specific CQA concentrations at the pools of each unit operation and drug substance were compared to the measured CQA distribution of LS runs. The OOS chance for the IPM was calculated by simply counting the number of simulations that are above the upper specification limit and dividing by the number of simulations. For the calculation of the OOS chance using the 9 large scale runs, a normal distribution was fitted to the data.

Figure 3, Figure 4, Figure 5 and Figure 6 show overlays of simulated and observed CQA distribution after each chromatography step for product related impurity 1 and 2, as well as process related impurity 1 and 2, respectively. For reasons of data security, all values have been normalized by the maximum observed or simulated CQA value. For the calculation of the observed distributions all 9 LS runs have been used and have been plotted. CQA distribution after chromatography column 3 (yellow colored bar in Figure 3, Figure 4, Figure 5 and Figure 6) can be regarded as drug substance since no further purification has been shown to occur at the ultrafiltration/diafiltration step.

From visual inspection the predicted distributions for each CQA nicely overlap with the observed distributions at each chromatography step. This is also reflected in good agreement of simulated and measured OOS probabilities at drug substance level, which are displayed in the title of each subfigure, except for process related impurity 2. Also, the skewness of the measured CQA distribution is well described by the model (e.g. positive skewness of the product related impurity 1 distribution at chromatography column 2 in Figure 3). Herein, we regard the model as valid for further investigations such as varying set-point conditions or accelerated variance of PPs.

For process related impurity 2, the variance of the predicted specific CQA concentrations is larger than the observed variance, especially at chromatography column 3 level, as shown in Figure 6. However, the mean prediction at chromatography column 3 level is very close to the observed runs. The simulated OOS events of the IPM are 9.1 % whereas only 0 % when calculating from LS data. This gap in predicted versus observed OOS events might be caused by an different mean response of the scale down model at set-point conditions, which was used to conduct the experiments, an overlooked effect of a PP onto this CQA, an overlooked spiking model, or the gap is introduced by the selection of the

current large scale runs which show a too low OOS chance. For the first two issues power analysis for the insignificant models terms needs to identify if additional experiments need to be conducted to make sure that no critical effect has been overlooked [9]. Whereas, the latter possibility indicates a risk that was uncovered by the IPM and has luckily not been observed during LS runs. Herein, counter actions might be taken such as an increase of specific purification capacity in primary recovery.

For product related impurity 2, the OOS chances for the IPM and the observed data are equally around 7 % as shown in Figure 4. Since for this CQA two statistical models as a function of PPs at chromatography column 1 and chromatography column 3 could be established (Table 2), parameter sensitivity analysis using the IPM can reveal optimization potential to increase process robustness for this CQA.



**Figure 3.** Comparison of simulated (top) product related impurity 1 distribution and observed (bottom) product related impurity 1 from LS after each column step. Normalization was performed by diving by the maximum observed $c_{CQA}$.
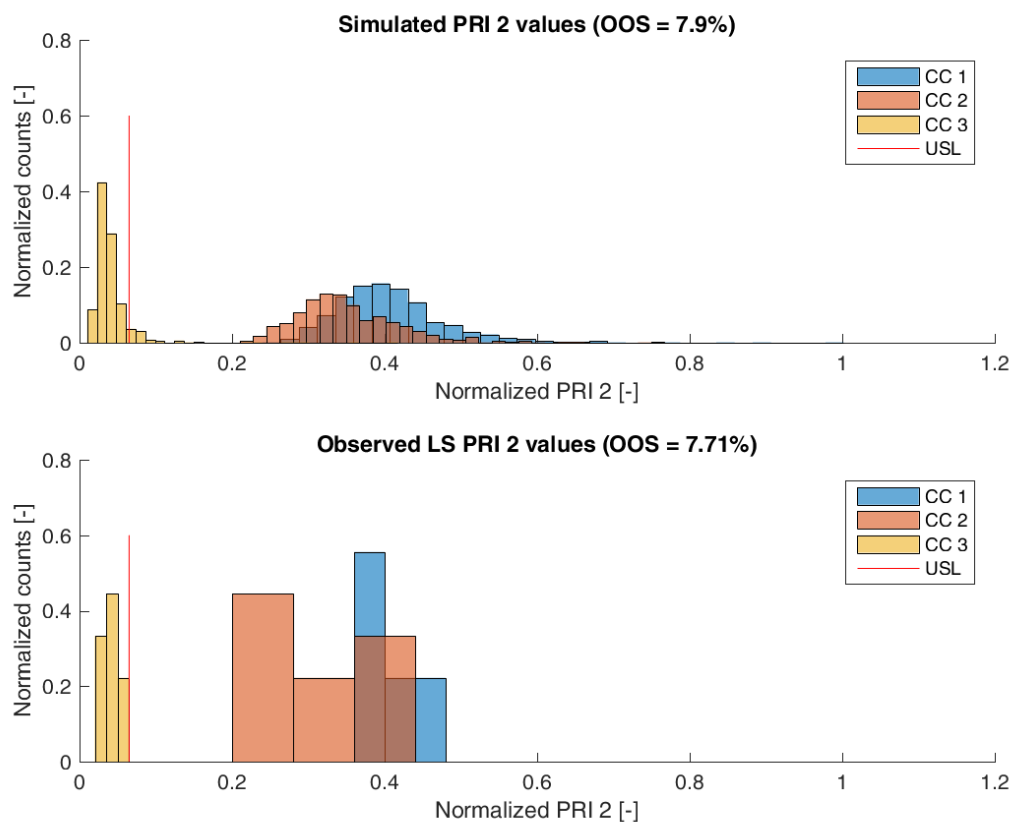
**Figure 4.** Comparison of simulated (top) product related impurity 2 distribution and observed (bottom) product related impurity 2 from LS after each column step. Normalization was performed by diving by the maximum observed $c_{CQA}$.
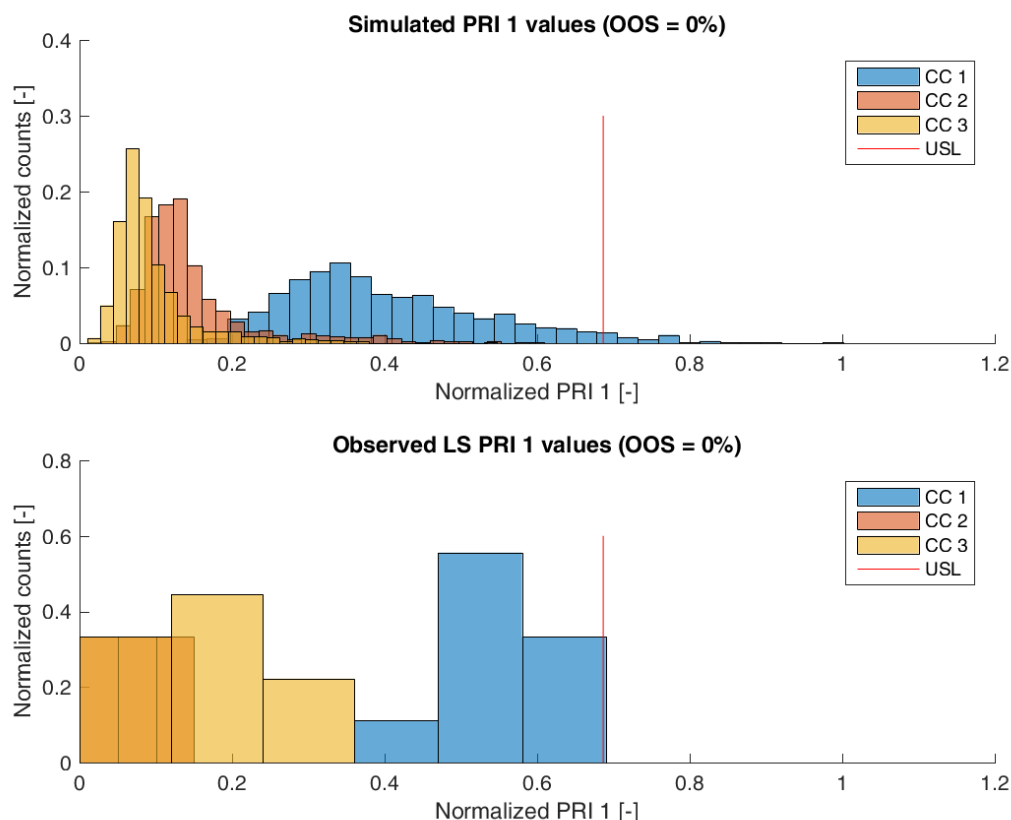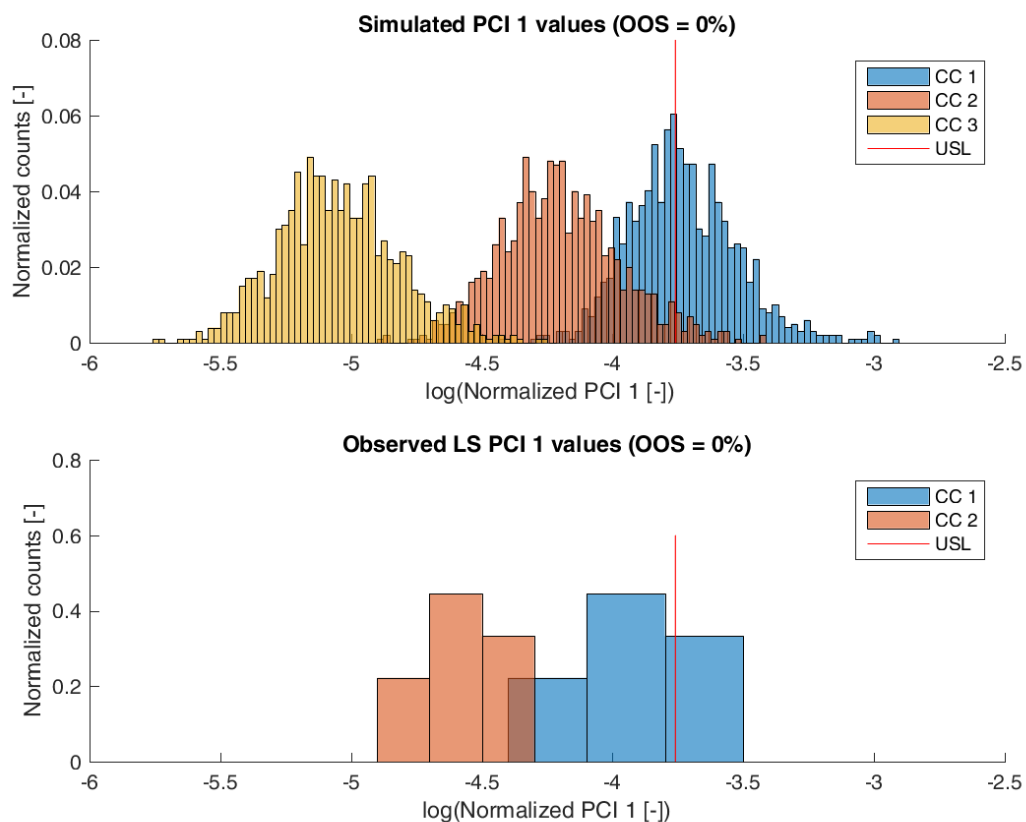
**Figure 5.** Comparison of simulated (top) process related impurity 1 distribution and observed (bottom) process related impurity 1 from LS after each column step. For chromatography column 3 pool, no process related impurity 1 value was observed above LoQ, therefore, no histogram bar is plotted for the observed values at chromatography column 3 pool. Normalization was performed by diving by the maximum observed $c_{CQA}$.

**Figure 6.** Comparison of simulated (top) process related impurity 2 distribution and observed (bottom) process related impurity 2 from LS after each column step. Normalization was performed by diving by the maximum observed $c_{CQA}$.
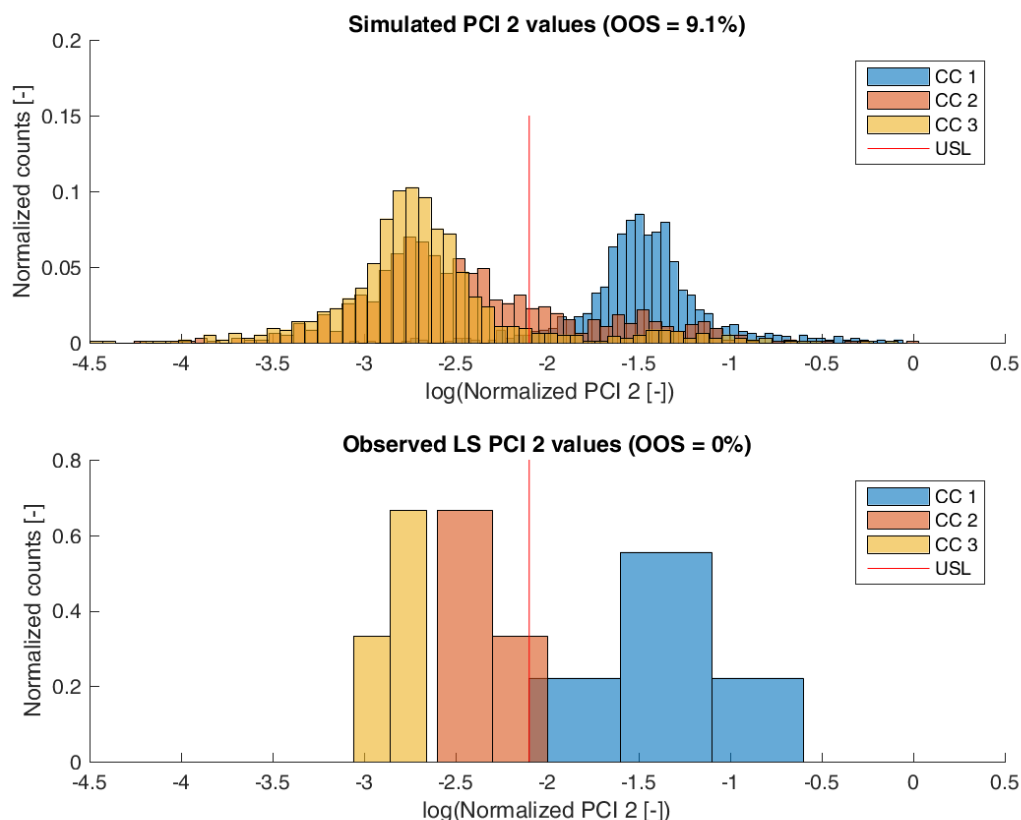
### 3.3. Impact of accelerated variation in process parameters on drug substance

Parameter sensitivity analysis (PSA) was performed to assess how a change in set-point or variance of controlled PPs influences OOS events at drug substance. PSA was conducted as follows: Each PP was varied individually regarding its mean and variance and resulting change in OOS events was measured. If interaction effects of parameters have been detected within DoE models, those parameters can be varied simultaneously to study this effect. However, this was not the case for any model established in this study. Moreover, since the model was built only on a segment of all unit operations, we are interested in how an altered performance of the fermentation and primary recovery – leading to an increased impurity burden at the load of chromatography column 1 – will impact on drug substance. Therefore, the specific impurity concentration at the loading of the chromatography column 1 was also varied in a parameter sensitivity analysis.

Results of an example of such an analysis are shown for product related impurity 2 (Figure 7), where in panel A the change of OOS events as a function of change in percent of set-point settings of all process parameters is displayed. As can be seen from this subfigure, only a change in pH and wash strength of chromatography column 1 leads to a drastic change in OOS events. This is expected since both factors are part of the DoE model (see Table S1 of supplementary materials in section 9.4). In more detail, both factors have a favorable direction in terms of reduction of OOS events (lowered pH and increased wash strength). For example, a reduction of the pH value by 10 % of the set-point leads to a reduction of OOS events from 7 % to around 3 %. Interestingly, a change in variance of those two process parameters by ± 50 % is not impacting at the OOS events (Figure 7B). This sounds contradictory at first glance, however, since a variance increase to a certain extent will also drive a lot of simulations to the more favorable side (lowered pH and increased wash strength), the overall OOS chance remains similar to the initial estimate. This also emphasizes the well-known fact that optimization should be rather addressed via a change in the set-point than reduction of variance, which is in general even harder to

accomplish. A change in initial product related impurity 2 burden after primary recovery propagates as well into drug substance, which can be explained by the fact that no spiking model could be established for this CQA at any unit operation as shown in Figure 7C. In detail, a 10 % reduction of the specific product related impurity 2 concentration after primary recovery lowers the OOS events by another 3 %. Therefore it would be favorable to lower the pH of chromatography column 1 and reduce the impurity burden already after primary recovery using prior knowledge or build models that capture the interaction of fermentation- and primary recovery parameters on this CQA. Thereby OOS events could be lowered for product related impurity 2 down to 1 % or less. In order not to increase the OOS probability for another CQA by changing those two process parameters, one would need to consider also their impact onto the residual CQAs. This is not shown here since we only wanted to introduce the methodology for a potential application of the IPM and due to reasons of simplicity.
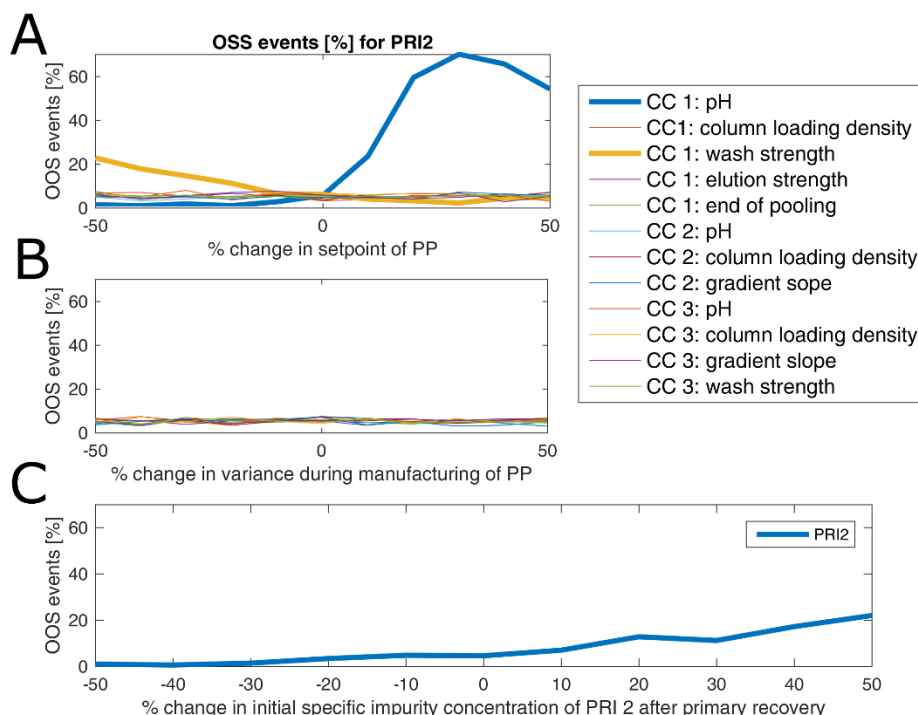


**Figure 7**. Estimated OOS event for product related impurity 2 at drug substance as a function of change in set-point (A) and variance (B) of all PPs as well as a function of increased specific impurity concentration after primary recovery (C). Deviations in set-point of pH and salt concentration in wash of chromatography column 1 impact severely on OOS chance, which is not the case when variance in PPs increases by up to 50%. A change of specific product related impurity 2 concentration at the primary recovery level will increase OOS chances, too.

## 4. Conclusions

Here we have shown how in using an IPM it was possible to demonstrate that sufficient process knowledge is available from process development to describe impurity clearance of process related impurity 1 and 2, as well as product related impurity 1 and 2. The distributions of simulated and observed CQAs are in good agreement to each other and make it possible to quantify the risk of not meeting product specifications under normal operating conditions, something which is often not possible due to limited large scale runs.

For product related impurity 1 and process related impurity 1, both, the predicted OOS chance by the IPM as well as the observed OOS chance is numerically close to 0 %. Herein, the process design can be validated in respect to those CQAs. In a first application of the IPM within a parameter sensitivity approach, it was possible to identify potential changes in process parameter set-points that will potentially decrease the chance of OOS events for the product related impurity 2 from 7 % to 1 %. For process related impurity 2 the mean prediction of clearance within the IPM is similar to that obtained

from LS measurements, however, the model predicts a 9.1 % chance to be above drug substance specification, whereas current large scale data estimate 0 % OOS chance. Since no statistical model could be established that might be used for optimization, process changes might be introduced. Here, IPM can be used within a model life-cycle approach as an enabler in change management. In case parts or entire unit operations are exchanged or included into an existing process design the IPM can predict the mutual performance of this change in the context of existing clearance capacity. This can be achieved by replacement with statistical models of respective unit operations. Thereby, overall performance of the changed process design can be assessed in terms of OOS events.

Furthermore, it should be emphasized that this model, in accordance with current opinion, is not finished in the traditional sense, but is expected to incorporate any future experiments and GMP runs for model refinement and application in further PV stages. Thereby it is expected that new or insufficiently studied dependencies between PPs and CQAs can be incorporated as identified.

## References

1. U.S. Department of Health and Human Services Process Validation: General Principles and Practices - UCM070336.pdf. **2011**.
2. Guideline, I. H. T. Pharmaceutical Development Q8 (R2). *Curr. Step* **2009**, *4*.
3. Guideline, I. H. T. Quality risk management. *Q9 Curr. Step* **2005**, *4*, 408.
4. Katz, P.; Campbell, C. FDA 2011 process validation guidance: process validation revisited. *J. GXP Compliance* **2012**, *16*, 18.
5. Peterson, J. J.; Lief, K. The ICH Q8 Definition of Design Space: A Comparison of the Overlapping Means and the Bayesian Predictive Approaches. *Stat. Biopharm. Res.* **2010**, *2*, 249–259, doi:10.1198/sbr.2009.08065.
6. Herwig, C.; Wölbeling, C.; Zimmer, T. A Holistic Approach to Production Control. *Pharm. Eng.* **2017**.
7. Bonate, P. L. A brief introduction to Monte Carlo simulation. *Clin. Pharmacokinet.* **2001**, *40*, 15–22.
8. Goudar, C. T.; Biener, R.; Konstantinov, K. B.; Piret, J. M. Error propagation from prime variables into specific rates and metabolic fluxes for mammalian cells in perfusion culture. *Biotechnol. Prog.* **2009**, *25*, 986–998, doi:10.1002/btpr.155.
9. Iman, R. L. Latin hypercube sampling. *Encycl. Quant. Risk Anal. Assess.* **2008**.
10. Singhee, A.; Rutenbar, R. A. Why Quasi-Monte Carlo is Better Than Monte Carlo or Latin Hypercube Sampling for Statistical Circuit Analysis. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2010**, *29*, 1763–1776, doi:10.1109/TCAD.2010.2062750.
11. Zahel, T.; Marschall, L.; Abad, S.; Vasilieva, E.; Maurer, D.; Mueller, E. M.; Natschläger, T.; Brochard, C.; Reinisch, D.; Herwig, C. Criticality Assessment Workflow for Biopharmaceutical Process Validation Stage 1. *Manuscr. Preperation*.

# 6 Conclusion

The aim of this work is to enrich traditional manufacturing PV stage 1 workflows with data science tools that aid in the scientific risk estimation and mitigation to ultimately decrease process variance and patient risks.

In this work it was possible to establish an advanced workflow for biopharmaceutical manufacturing process validation stage 1 that improves process robustness and decreases patient risk. Three areas of improvement to traditional manufacturing process validation have been presented:

- Fermentation is one of the most complex unit operations and thereby also biopharmaceutical manufacturing differentiates to small molecule production. CQAs are analytically hard to measure at that early stage of the process. Therefore, indirect measures by means of physiological activity and specific turnover rates are highly valuable. However, only by knowing accuracy of that measures makes it possible to estimate the risk associated with decisions based on that quantities. Therefore, we developed methods to derive specific rates that show constant signal to noise ratio derived from online and offline raw measurements. Using that methods it is possible to obtain a fair and well-balances comparison of specific rates across scales. This is essential for SDM qualification but they may also serve as a high-accuracy response for experimental evaluation of parameter criticality.

- Less focus has been given to the unknown risk of a process in traditional process validation stage 1 approaches. How likely is the chance that we missed to detect a critical influencer that will show up in later stages of production? This kind of questions have not been targeted so far using data driven methods, also due to missing statistical approaches. Those have been presented in this work and their benefits in mitigating "unknown unknown" CPPs to "known unknown" CPPs has been demonstrated successfully. Moreover, it has been shown how a control strategy can be set up for those "known unknown" CPPs. This methodology, is not limited or special for biopharmaceutical process validation but can be applied to all areas of manufacturing where unknown risk is inferred from experimental data.

- Another typical aspect of biopharmaceutical processes is the usage of multiple unit operations that are carried out sequentially for production and clearance of to the final product. In this setting understanding of interactions between critical parameters of single unit operations is key for a holistic risk estimation. The presented integrated process model is connecting statistical regression models of single unit operations together and predicting out of specification probabilities as a function of CPPs.

However, only by making sure that no CPP has been overlooked during task 5 of the presented workflow it is possible to correctly estimate process robustness. Thereby the overall risk of a change in a single CPP can be estimated and control strategies might be adapted accordingly. Thereby, sufficient confidence can be obtained that the process will be capable of consistently delivering quality product in future.

The overall presented workflow consists of sequence of steps that might be followed in the presented order when applying the stage 1 for the first time to a specific product. There are clear inputs and outputs for each step that need to be in place as described in this work, however, they might have been achieved earlier in the product life cycle. If single steps are skipped without having in enough confidence in the achieved output, serious impact can be expected ultimately leading to increased process variance. For example, by skipping data mining or not having a full understanding of the meaning and reliability of the manufacturing data, it will not be possible to perform a scientifically sound risk assessment nor will it be possible to judge the predictability of the SDM. Since risk assessment (task 3) and SDM qualification (task 4) do not rely on input of each other directly they might be performed in parallel or independently from each other.

Overall, the presented workflows enables to scientifically sound estimate the risk on product quality associated with certain PPs or unit operations. Measures of that risk are acceptance criteria, power levels of PPs to be overlooked, similarity of SDM as well as holistic criticality of single PPs. Those estimates need to be understood as tentative measures which are constantly refined during process validation life cycle. As soon as more data from manufacturing runs is collected a periodic update of those measures is useful and regulatory requested to constantly demonstrate risk awareness.

Due to the complexity of the tasks, they need to be addressed in a multidisciplinary team formed by statisticians, process experts, quality control, analytical chemists and validation expert. This is also recommended in the FDA process validation guideline [6].

This work has proven that the presented novel methods can increase the science based process knowledge, risk awareness and how to impose an appropriate control strategy for biopharmaceutical processes. Moreover, those methods have proven to give evidence "that a process is capable of consistently delivering quality product", which is literally the goal of process validation stage 1 [6].

# 7 Outlook

In contrast to small molecule production, biopharmaceutical manufacturing differs also by the amount of redundant analytical methods employed to characterize the product. A bunch of analytical methods are used to characterize complex protein or nucleic acid structure, its isoforms and impurities. Therefore, we are rich in redundant information about the quality of our product, which we should make use of. Multivariate statistics are an excellent tool to use redundant variables and estimate latent variables that can be compared between scales. Therefore, multivariate equivalence testing offers a huge potential to increase accuracy of SDM qualification without increasing the number of experiments or analytical measurements required. Moreover, this can be enriched by usage of Bayesian methods to incorporate physical constraints and prior knowledge and thereby additional elevate power of the detecting equivalence or a reduction of experimental effort.

A hurdle in applying equivalence testing for SDM qualification is the setting of practical relevant levels of difference between manufacturing and small scale, EACs. This is potentially also the reason why still less scientifically sound methods such as checking if all runs are within 3 sigma of the large scale runs are applied. Strategies that try to estimate the EACs with a constant times the standard deviation of the manufacturing/reference runs (i.e. $EAC = f \cdot \sigma_{reference}$) always award for large scattering in the manufacturing/reference group. This might be applicable when a large number of replicate runs are available that have proven to deliver high product quality. However, in stage 1 of process validation only few manufacturing runs are available, which limits the estimation of the true scattering. Here I want to briefly give possibilities how EACs can be achieved systematically for all unit operations:

- For responses which are use across multiple processes and products, such as viable cell density for fermentation processes or product yields of chromatography columns, general EACs can be established by applying a questionnaire within process experts of the firm. Questions should aim to answer what is the difference we are likely to accept as no practical relevant difference. A mean value of all those answers could be used as a general standard.

- For responses which are product specific or less prior knowledge is available:
  - If acceptance criteria are available the shift in large scale population that still leads to an acceptable process capability index (e.g. Ppk > 1) can be regarded as practical relevant difference to impact product quality. In some cases the herein derived EACs might be tightened in case of extremely robust process steps with extraordinary high process capability. Otherwise, this may lead to non-informatively wide EACs.

- If failed batches are available that can be regarded as non-similar to the existing process, a decision boundary between acceptable batches, known to have reached specifications and those failed batches can be established. This can be performed using classification algorithms such as support vector machines where the decision boundary is a balanced margin between similar and non-similar observations.

However, in cases where only a limited amount of manufacturing runs exist, only tentative estimates of the EAC can be established, which need to be refined during the process validation life cycle.

# 8 References

1. Guideline, I. H. T. Quality risk management. *Q9 Curr. Step* **2005**, *4*, 408.
2. Katz, P.; Campbell, C. FDA 2011 process validation guidance: process validation revisited. *J. GXP Compliance* **2012**, *16*, 18.
3. Aleem, H.; Zhao, Y.; Lord, S.; McCarthy, T.; Sharratt, P. Pharmaceutical process validation: An overview. *Proc. Inst. Mech. Eng. Part E J. Process Mech. Eng.* **2003**, *217*, 141–151, doi:10.1243/095440803766612801.
4. Ahir, K. B.; Singh, K. D.; Yadav, S. P.; Patel, H. S.; Poyahari, C. B. Overview of Validation and Basic Concepts of Process Validation. *Sch. Acad. J. Pharm.* **2014**, *3*, 178–90.
5. Guideline, I. H. T. Pharmaceutical Development Q8 (R2). *Curr. Step* **2009**, *4*.
6. FDA Guidance for Industry Available online: https://www.fda.gov/downloads/drugs/guidances/ucm070336.pdf (accessed on Dec 4, 2017).
7. Seely, J. E.; Seely, R. J. A Rational, Step-Wise Approach to Process Characterization Available online: http://www.biopharminternational.com/rational-step-wise-approach-process-characterization (accessed on Nov 9, 2015).
8. Rodriguez, J. *CAPA in the pharmaceutical and biotech industries: how to implement an effective nine step program*; Woodhead Publishing series in biomedicine; WP, Woodhead Publishing, an imprint of Elsevier: Amsterdam Boston Cambridge, 2016; ISBN 978-1-907568-58-9.
9. Walsh, G. Biopharmaceuticals and biotechnology medicines: an issue of nomenclature. *Eur. J. Pharm. Sci.* **2002**, *15*, 135–138.
10. *Quality by design for biopharmaceuticals: principles and case studies*; Rathore, A. S., Mhatre, R., Eds.; Wiley: Hoboken, N.J, 2009; ISBN 978-0-470-28233-5.
11. 12th Annual Report and Survey of  Biopharmaceutical Manufacturing  Capacity and Production Available online: http://www.bioplanassociates.com/publications/12th_Biomfg_Table_of_Content.pdf (accessed on Oct 23, 2017).
12. Rathore, A. S.; Sofer, G. *Process Validation in Manufacturing of Biopharmaceuticals, Third Edition*; CRC Press, 2012; ISBN 978-1-4398-5093-0.
13. Dahiya, S.; Khar, R. K.; Chhikara, A. Opportunities, challenges and benefits of using HACCP as a quality risk management tool in the pharmaceutical industry. *Qual. Assur. J.* **2009**, *12*, 95–104, doi:10.1002/qaj.446.
14. Zimmermann, H. F.; Hentschel, N. Proposal on How To Conduct a Biopharmaceutical Process Failure Mode and Effect Analysis (FMEA) as a Risk Assessment Tool. *PDA J. Pharm. Sci. Technol.* **2011**, *65*, 506–512, doi:10.5731/pdajpst.2011.00784.
15. Mollah, A. H. Application of failure mode and effect analysis (FMEA) for process risk assessment. *BioProcess Int.* **2005**, *3*, 12–20.

16. Schmidt, F. R. Optimization and scale up of industrial fermentation processes. *Appl. Microbiol. Biotechnol.* **2005**, *68*, 425–435, doi:10.1007/s00253-005-0003-0.

17. Takors, R. Scale-up of microbial processes: Impacts, tools and open questions. *J. Biotechnol.* **2012**, *160*, 3–9, doi:10.1016/j.jbiotec.2011.12.010.

18. Garcia-Ochoa, F.; Gomez, E. Bioreactor scale-up and oxygen transfer rate in microbial processes: An overview. *Biotechnol. Adv.* **2009**, *27*, 153–176, doi:10.1016/j.biotechadv.2008.10.006.

19. Shimoni, Y.; Goudar, C.; Jenne, M. Qualification of Scale-Down Bioreactors: Validation of Process Changes in Commercial Production of Animal-Cell-Derived Products, Part 1—Concept. *BioProcess Int* **2014**, *12*.

20. ICH Q11 Available online: https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q11/Q11_Step_4.pdf (accessed on Oct 25, 2017).

21. What is a (qualified) bioprocess scale down model? · EXPUTEC Available online: https://exputec.com/what-is-a-qualified-bioprocess-scale-down-model/ (accessed on Dec 4, 2017).

22. Berger, R. L.; Hsu, J. C.; others Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Stat. Sci.* **1996**, *11*, 283–319.

23. Limentani, G. B.; Ringo, M. C.; Ye, F.; Berquist, M. L.; McSorley, E. O. Beyond the t-test: statistical equivalence testing. *Anal. Chem.* **2005**, *77*, 221A–226A.

24. Hoffelder, T.; Gössl, R.; Wellek, S. Multivariate Equivalence Tests for Use in Pharmaceutical Development. *J. Biopharm. Stat.* **2015**, *25*, 417–437, doi:10.1080/10543406.2014.920344.

25. Chervoneva, I.; Hyslop, T.; Hauck, W. W. A multivariate test for population bioequivalence. *Stat. Med.* **2007**, *26*, 1208–1223, doi:10.1002/sim.2605.

26. Burdick, R. K.; LeBlond, D. J.; Pfahler, L. B.; Quiroz, J.; Sidor, L.; Vukovinsky, K.; Zhang, L. Process Design: Stage 1 of the FDA Process Validation Guidance. In *Statistical Applications for Chemistry, Manufacturing and Controls (CMC) in the Pharmaceutical Industry*; Springer International Publishing: Cham, 2017; pp. 115–154 ISBN 978-3-319-50184-0.

27. EMEA Guideline on Comparability of Biotechnology-Derived Medicinal Products Afer a Change in the Manufacturing Process.

28. ICH ICH Q5E: Comparability of Biotechnological/Biological Products Subject to Changes in Their Manufacturing Process.

29. FDA FDA Comparability Protocols for Human drugs and biologics.pdf.

30. Reflection paper on statistical methodology for the comparative assessment of quality attributes in drug development - Draft Available online: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/03/WC500224995.pdf (accessed on Oct 30, 2017).

31. Sagmeister, P.; Wechselberger, P.; Herwig, C. Information Processing: Rate-Based Investigation of Cell Physiological Changes along Design Space Development. *PDA J. Pharm. Sci. Technol.* **2012**, *66*, 526–541, doi:10.5731/pdajpst.2012.00889.

32. Jenzsch, M.; Gnoth, S.; Kleinschmidt, M.; Simutis, R.; Lübbert, A. Improving the batch-to-batch reproducibility of microbial cultures during recombinant protein production by regulation of the total carbon dioxide production. *J. Biotechnol.* **2007**, *128*, 858–867, doi:10.1016/j.jbiotec.2006.12.022.

33. Doran, P. M. *Bioprocess engineering principles*; 2nd ed.; Elsevier/Academic Press: Amsterdam ; Boston, 2013; ISBN 978-0-12-220851-5.

34. Van der Heijden, R.; Heijnen, J. J.; Hellinga, C.; Romein, B.; Luyben, Kc. Linear constraint relations in biochemical reaction systems: I. Classification of the calculability and the balanceability of conversion rates. *Biotechnol. Bioeng.* **1994**, *43*, 3–10.

35. Van der Heijden, R.; Romein, B.; Heijnen, J. J.; Hellinga, C.; Luyben, Kc. Linear constraint relations in biochemical reaction systems: II. Diagnosis and estimation of gross errors. *Biotechnol. Bioeng.* **1994**, *43*, 11–20.

36. Sagmeister, P.; Langemann, T.; Wechselberger, P.; Meitz, A.; Herwig, C. A dynamic method for the investigation of induced state metabolic capacities as a function of temperature. *Microb. Cell Factories* **2013**, *12*, 94.

37. Jobé, A. M.; Herwig, C.; Surzyn, M.; Walker, B.; Marison, I.; von Stockar, U. Generally applicable fed-batch culture concept based on the detection of metabolic state by on-line balancing. *Biotechnol. Bioeng.* **2003**, *82*, 627–639, doi:10.1002/bit.10610.

38. Zahel, T.; Sagmeister, P.; Suchocki, S.; Herwig, C. Accurate Information from Fermentation Processes – Optimal Rate Calculation by Dynamic Window Adaptation. *Chem. Ing. Tech.* **2016**, n/a-n/a, doi:10.1002/cite.201500085.

39. Steinwandter, V.; Zahel, T.; Sagmeister, P.; Herwig, C. Propagation of measurement accuracy to biomass soft-sensor estimation and control quality. *Anal. Bioanal. Chem.* **2016**, doi:10.1007/s00216-016-9711-9.

40. Zahel, T.; Marschall, L.; Abad, S.; Vasilieva, E.; Maurer, D.; Mueller, E. M.; Murphy, P.; Natschläger, T.; Brocard, C.; Reinisch, D.; Sagmeister, P.; Herwig, C. Workflow for Criticality Assessment Applied in Biopharmaceutical Process Validation Stage 1. *Bioengineering* **2017**, *4*, 85, doi:10.3390/bioengineering4040085.

41. Jones, B.; Nachtsheim, C. J. A class of three-level designs for definitive screening in the presence of second-order effects. *J. Qual. Technol.* **2011**, *43*, 1.

42. Tai, M.; Ly, A.; Leung, I.; Nayar, G. Efficient high-throughput biological process characterization: Definitive screening design with the Ambr250 bioreactor system. *Biotechnol. Prog.* **2015**, *31*, 1388–1395, doi:10.1002/btpr.2142.

43. Ke, W.; Ide, N. D.; Dirat, O.; Subashi, A. K.; Thomson, N.; Vukovinsky, K.; Watson, T. J. Statistical Tools to Aid in the Assessment of Critical Process Parameters | Pharmaceutical Technology Available online: http://www.pharmtech.com/statistical-tools-aid-assessment-critical-process-parameters?pageID=1 (accessed on Oct 25, 2017).

44. Cohen, J. *Statistical power analysis for the behavioral sciences*; Rev. ed.; Academic Press: New York, 1977; ISBN 978-0-12-179060-8.

45. Thomas, L.; Krebs, C. J. A review of statistical power analysis software. *Bull. Ecol. Soc. Am.* **1997**, *78*, 126–138.

46. Thomas, L. Retrospective Power Analysis. *Conserv. Biol.* **1997**, *11*, 276–280, doi:10.1046/j.1523-1739.1997.96102.x.

47. Peres-Neto, P. R.; Olden, J. D. Assessing the robustness of randomization tests: examples from behavioural studies. *Anim. Behav.* **2001**, *61*, 79–86, doi:10.1006/anbe.2000.1576.

48. Characterizing Unknown Unknowns Available online: https://www.pmi.org/learning/library/characterizing-unknown-unknowns-6077 (accessed on Oct 25, 2017).

49. Ponton, J. W. Dynamic process simulation using flowsheet structure. *Comput. Chem. Eng.* **1983**, *7*, 13–17.

50. Pattison, R. C.; Baldea, M. Equation-oriented flowsheet simulation and optimization using pseudo-transient models. *AIChE J.* **2014**, *60*, 4104–4123, doi:10.1002/aic.14567.

51. *Product and process design principles: synthesis, analysis, and evaluation*; Seider, W. D., Seider, W. D., Eds.; 3rd ed.; Wiley: Hoboken, NJ, 2009; ISBN 978-0-470-04895-5.

52. Sen, M.; Chaudhury, A.; Singh, R.; John, J.; Ramachandran, R. Multi-scale flowsheet simulation of an integrated continuous purification–downstream pharmaceutical manufacturing process. *Int. J. Pharm.* **2013**, *445*, 29–38, doi:10.1016/j.ijpharm.2013.01.054.

53. Boukouvala, F.; Ramachandran, R.; Vanarase, A.; Muzzio, F. J.; Ierapetritou, M. G. Computer aided design and analysis of continuous pharmaceutical manufacturing processes. *Comput. Aided Chem. Eng.* **2011**, *29*, 216–220.

54. Seely, R. J.; Munyakazi, L.; Haury, J. Statistical tools for setting in-process acceptance criteria. *Dev. Biol.* **2003**, *113*, 17–25.

55. Wang, X.; Germansderfer, A.; Harms, J.; Rathore, A. S. Using Statistical Analysis for Setting Process Validation Acceptance Criteria for Biotech Products. *Biotechnol. Prog.* **2007**, *23*, 55–60, doi:10.1021/bp060359c.

56. Harms, J.; Wang, X.; Kim, T.; Yang, X.; Rathore, A. S. Defining Process Design Space for Biotech Products: Case Study of Pichia pastoris Fermentation. *Biotechnol. Prog.* **2008**, *24*, 655–662, doi:10.1021/bp070338y.

57. Peterson, J. J.; Lief, K. The ICH Q8 Definition of Design Space: A Comparison of the Overlapping Means and the Bayesian Predictive Approaches. *Stat. Biopharm. Res.* **2010**, *2*, 249–259, doi:10.1198/sbr.2009.08065.

58. Zahel, T.; Hauer, S.; Mueller, E. M.; Murphy, P.; Abad, S.; Vasilieva, E.; Maurer, D.; Brocard, C.; Reinisch, D.; Sagmeister, P.; Herwig, C. Integrated Process Modeling—A Process Validation Life Cycle Companion. *Bioengineering* **2017**, *4*, 86, doi:10.3390/bioengineering4040086.

59. Bonate, P. L. A brief introduction to Monte Carlo simulation. *Clin. Pharmacokinet.* **2001**, *40*, 15–22.

60. Skrondal, A. Design and Analysis of Monte Carlo Experiments: Attacking the Conventional Wisdom. *Multivar. Behav. Res.* **2000**, *35*, 137–167, doi:10.1207/S15327906MBR3502_1.

# 9 Annex: Supporting Information to Manuscripts

## 9.1 Supporting Information: Accurate Information from Fermentation Processes - Optimal Rate Calculation by Dynamic Window Adaptation"

## S1 Mechanistic model formulation for generation of model based in silico data

For the generation of in silico data, which is used to derive an optimal rate calculation workflow for microbial fed-batch processes, a mechanistic model is applied as briefly described in section 2.1 of the main article. Here we want to give a more detailed view on this model.

An industrial fed-batch feeding profile according to Wechselberger [1] was simulated with a short exponential feeding phase followed by a constant feeding phase as investigated previously in literature (Figure 4 of main article). A decrease of the biomass yield coefficient with process maturity during the constant feeding, as frequently observed in industrial processes due the effects of metabolic load [2], was included in the model. This increases the signal dynamics of the resulting biomass formation rate and thereby a considerable and industrially relevant benchmark signal ($r_x$) for testing different rate calculation algorithms is obtained.

## S2 Formulation of extended Kalman filter configuration

In the following, the basic formulation of the Kalman filter configuration, as referred to from section 2.4 of the main article, is described in detail:

The dynamic model:

$$\dot{x}(t) = f\big(x(t)\big) + n(t) \text{ (1)}$$

where $x(t)$ is the state vector and $n(t)$ has the covariance matrix $Q(t)$.

Observation model:

$$l(t) = H \cdot x(t) + w(t) \text{ (2)}$$

Where $l(t)$ is the vector of observations and the covariance matrix of $w(t)$ is $R(t_i)$. In our case it was approximated that there is no transfer matrix for the measurement and therefore $H$ was set to $I$.

The a priori state:

$$\dot{x}^-(t) = f\big(x^-(t)\big) \text{ (3)}$$

The state transition matrix ($\Phi$) can obtained by:

$$\frac{d}{dt} \boldsymbol{\Phi}_{t_{i-1}}^{t_i} = \boldsymbol{F}(t_i) \cdot \boldsymbol{\Phi}_{t_{i-1}}^{t_i} \ (4)$$

And the dynamic matrix ($\boldsymbol{F}(t_i)$) can be approximated with a Tayler series with respect to $x$ at $t_i$ terminated after the first term:

$$\boldsymbol{F}(t_i) = \left. \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \right|_{\boldsymbol{x} = \boldsymbol{x}^-(t_i)} \ (5)$$

The a priori covariance matrix $\boldsymbol{P}^-(t_i)$:

$$\boldsymbol{P}^-(t_i) = \boldsymbol{\Phi}_{t_{i-1}}^{t_i} \cdot \boldsymbol{P}^+(t_{i-1}) \cdot (\boldsymbol{\Phi}_{t_{i-1}}^{t_i})^T + \int_{t_{i-1}}^{t_i} \boldsymbol{Q}(t) dt \ (6)$$

And $\boldsymbol{Q}(t)$ was set to zero since no process variance was expected for the in silico generated data.

Gain Matrix $\boldsymbol{K}(t_i)$ at time point $t_i$ since $\boldsymbol{H} = \boldsymbol{I}$:

$$\boldsymbol{K}(t_i) = \boldsymbol{P}^-(t_i) \cdot \left( \boldsymbol{P}^-(t_i) + \boldsymbol{R}(t_i) \right)^{-1} \ (7)$$

A posteriori state:

$$\boldsymbol{x}^+(t_i) = \boldsymbol{x}^-(t_i) + \boldsymbol{K} \cdot (\boldsymbol{l}(t_i) - \boldsymbol{x}^-) \ (8)$$

A posteriori covariance matrix:

$$\boldsymbol{P}^+(t_i) = (\boldsymbol{I} - \boldsymbol{K}(t_i)) \cdot \boldsymbol{P}^-(t_i) \ (9)$$

# S3 Derivation of the signal to noise ratio of rates using Gaussian error propagation

In order to obtain specific rates with a constant signal to noise ratio, error propagation from the original signals to the derived turnover rates needs to be investigated, as outlined briefly in section 3.1.1 of the main article.

The influence of the absolute measurement uncertainty boundary ($U_y$) of the signal $y$ onto a derived signal $r$ can be approximated using a Taylor expansion [3]:

$$r(y + U_y) = r(y) + \frac{1}{1!} \frac{dr(y)}{dy} \cdot U_y + \frac{1}{2!} \frac{d^2 r(y)}{dy^2} \cdot (U_y)^2 + \cdots \ (10)$$

For an approximate solution the Taylor expansion can be terminated after the second term and the resulting absolute deviation of the derived signal ($\Delta U_r$) can be written as:

$$r(y + U_y) - r(y) = U_r = \frac{dr(y)}{dy} \cdot U_y \quad (11)$$

If the derived signal depends on more than one input variable, the input variables are uncorrelated and the uncertainty of the input signal is only known by its boundaries, which is the typical case for biotechnological rate calculations, we can write in analogy:

$$U_r = \left| \frac{\partial r}{\partial y_1} \right| \cdot U_{y_1} + \left| \frac{\partial r}{\partial y_2} \right| \cdot U_{y_2} \quad (12)$$

The finite difference approximation is used to obtain the derivate of the signal $y$:

$$r = \frac{y_2 - y_1}{\Delta t} \quad (13)$$

By inserting Eq. 13 in Eq. 12 we can finally estimate the absolute uncertainty of the derived rate, which is:

$$U_r = \frac{U_{y_1} + U_{y_2}}{\Delta t} \quad (14A)$$

If the uncertainty of the input signals is known by its standard deviation we can alternatively formulate the simple rate uncertainty, $u_r$:

$$u_r = \frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\Delta t} \quad (14B)$$

In this contribution we use Eq. 14A to derive the SNR of the rate (Eq. 15). Using the obtained SNR of the rate (Eq. 15) the time window can be dynamically widened, where the resulting SNR on the rate is too low, yielding to significant rates with predefined SNR at highest possible time resolution.

$$SNR_r = \frac{r}{U_r} \quad (15)$$

# S4 Rational driven parameter selection for existing pre- and post-processing methods

Before using filtering and smoothing algorithms, the selection of algorithm parameters has to be performed carefully on a rational basis in order not to deteriorate the signal characteristics, as described in the section 3.1.2 of the main article. However, so far the determination of cut-off frequencies and smoothing windows remain an empirical endeavour and no rationales are

provided [4,5]. Here, we want to present some bioprocess rationales supporting a rationally sound filter design.

The cut-off frequency of the Butterworth filter for post-processing rates can be derived from a Fourier analysis of the expected rate. For example, in a carbon limited fed-batch the change in the substrate uptake rate cannot be faster than the change in the feeding rate assuming metabolic shifts are slower. Therefore, a Fourier analysis was performed on the supposed feeding rate and a cut-off frequency of 0.5 mHz was determined for the in silico data set presented in section 2.1 of the publication, since no higher frequency components were present. This cut-off frequency can also be used for the real process data. Here, the signal dynamics of the feeding rate were even less pronounced since no exponential feed was applied but only constant and linear feeding rates. This procedure has to be evaluated for each rate separately. Moreover, the cut-off frequency cannot be rationally derived for original raw signals (e.g. scale signal, pump signal) and therefore all high frequency components are allowed. This can be easily illustrated by the fact that a pump may be switched on/off instantaneously but the resulting change in the specific substrate uptake will occur smoothly. Therefore, filtering using the frequency domain is suggested only for physiological rates. All presented Butterworth filters are $4^{th}$-order low-pass filters, which corresponds to a moderate frequency roll off.

For the Savitzky-Golay smoothing of online signals (measurement period < 2min), which have a supposedly smooth deterministic signal compound, a first order polynomial is used. Furthermore, the window width should be smaller than the width of the fastest signal change but larger than the supposed width of spikes and ripples. In our case, for the scale signals, this was approximated to 20 data points at a sampling interval of 20 s.

# S5 Additional rate calculation workflows for industrial microbial fed-batch fermentation deriving $r_x$ and $r_S$ using the generic model-based algorithm

As outlined in the section 3.2.1 of the publication, the optimal rate calculation workflow for an industrial microbial fed-batch fermentation can be found using the generic model-based algorithm presented in section 3.1.2. In section 3.2.1 the most relevant pre-processing and rate calculation combinations are shown in order to derive volumetric biomass formation rate ($r_x$) and substrate uptake rate ($r_S$). In Table 1 and Table 2 we want to show more investigated combinations of pre-processing and rate calculation steps for deriving $r_S$ and $r_x$, respectively. The different combinations were evaluated in respect to their accuracy to the true rates using

the normalized root mean square error (NRMSE) and in respect to their precision using the mean relative error on the rate ($\Delta r_{rel}$).

*Table 1: comparison of different rate calculation algorithms for the substrate uptake rate ($r_S$) originating from a scale signal (online signal) by means of NRMSE to the true rate and relative errors of the rate ($\Delta r_{rel}$) at different absolute noise levels. Used abbreviations for algorithms: DWR5: rate calculation with dynamic window adaption with SNR 5 on resulting rates, FWR2: rate calculation with fixed window size of 2 points, FWR4: rate calculation with fixed window size of 4 points, SGF20: Savitzky-Golay smoothing with a window size of 20 data points, SGR20: Savitzky-Golay first derivative with a window size of 20 data points, BwF0.5: Butterworth low pass filter with 0.5 mHz cut-off frequency, n.a. refers to not available, since the first derivative with Savitzky-Golay algorithms does not provide estimates about the precision of the calculated rate.*

| Added absolute noise [g] | 0.05 | | 0.1 | |
|---|---|---|---|---|
| | NRMSE | $\Delta r_{rel}$ | NRMSE | $\Delta r_{rel}$ |
| DWR5 | 0.111 | 0.132 | 0.113 | 0.134 |
| SGF20 + DWR5 | 0.043 | 0.129 | 0.067 | 0.132 |
| SGF20 + DWR5 + BwF0.05 | 0.027 | 0.129 | 0.047 | 0.132 |
| FWR2 | 1.712 | 6.194 | 3.422 | 6.882 |
| SGF20 + FWR2 | 0.086 | 1.976 | 0.171 | 4.112 |
| FWR4 | 0.578 | 2.795 | 1.148 | 6.603 |
| SGF20 + FWR4 | 0.050 | 0.652 | 0.099 | 1.327 |
| SGF20 + FWR4 + BwF0.5 | 0.035 | 0.651 | 0.067 | 1.319 |
| SGR20 | 0.048 | n.a. | 0.094 | n.a. |
| SGR20 + BwF0.5 | 0.024 | n.a. | 0.048 | n.a. |

Table 2: Comparison of different rate calculation algorithms for the volumetric biomass formation rate ($r_x$) originating from an OD-measurement (offline signal) by means of NRMSE to the true rate and relative errors ($\Delta r_{rel}$) of the rate at different relative noise levels of the OD-measurement. Used abbreviations for algorithms: DWR5: rate calculation with dynamic window adaption with SNR of 5 on resulting rates, FWR2: rate calculation with fixed window size of 2 points, FWR3: rate calculation with fixed window size of 3 points, FWR4: rate calculation with fixed window size of 4 points, PBR: window size equals phase boundary, KaF: Kalman filtering.

| Added relative error to biomass signal [%]: | 0.5 - 1 | | 2 - 2.5 | | 3.5 - 4 | | 5 - 5.5 | |
|---|---|---|---|---|---|---|---|---|
| Applied algorithms: | NRMSE | $\Delta r_{rel}$ | NRMSE | $\Delta r_{rel}$ | NRMSE | $\Delta r_{rel}$ | NRMSE | $\Delta r_{rel}$ |
| DWR5 | 0.15 | 0.12 | 0.22 | 0.14 | 0.22 | 0.15 | 0.34 | 0.16 |
| FWR2 | 0.37 | 3.46 | 0.99 | 5.4 | 1.71 | 7.03 | 2.35 | 6.12 |
| FWR3 | 0.24 | 2.67 | 0.5 | 3.21 | 0.88 | 6.14 | 1.11 | 3.35 |
| FWR4 | 0.26 | 1 | 0.39 | 7.26 | 0.57 | 4.57 | 0.76 | 4.97 |
| PBR | 0.82 | 0.03 | 0.82 | 0.08 | 0.83 | 0.13 | 0.84 | 0.18 |
| KaF + DWR5 | 0.11 | 0.09 | 0.13 | 0.1 | 0.16 | 0.1 | 0.19 | 0.1 |
| KaF + FWR2 | 0.11 | 0.15 | 0.14 | 0.16 | 0.2 | 0.19 | 0.26 | 0.27 |

# S6 Proof of equivalence of finite difference approximation in combination with linear interpolation and "integral" calculation of specific rates (q$_i$)

In this contribution we presented an optimal rate calculation workflow that uses finite difference approximation for deriving volumetric rates. Dividing those volumetric rates by the concentration of biomass, specific rates for each species $i$ ($q_i$) can be obtained. However, a different specific rate calculation method for mammalian cell culture than for microbial fermentations is commonly applied, as discussed in section 4.4 of the main article. It was suggested to use the ratio of the observed mass difference during a time window ($t - t_0$) of species $i$ ($\Delta m_i$) to the integral of viable cells ($IVC$) according to Eq. 16 and 17 [6,7]. The numerical integration in Eq. 17 was performed using trapezoidal integration.

$$IVC = \int_{t0}^{t} \gamma_{X_V} \cdot dt \quad (16)$$

$$q_{i,I} = \frac{\Delta m_i}{\int_{t0}^{t} \gamma_{X_V} \cdot dt} = \frac{\Delta m_i}{\frac{1}{2}(\gamma_{X_{V,t}} + \gamma_{X_{V,t0}}) \cdot \Delta t} \quad (17)$$

It can be easily shown that the finite difference approximation delivers the same specific rate ($q_{i,FD}$) at $t_0 + \frac{t-t_0}{2}$ using linear interpolation for the viable cells concentration ($X_V$) as shown in Eq. 18:

$$q_{i,FD} = \frac{\Delta m_i}{\Delta t} \cdot \frac{1}{\gamma_{X_{V,t0+(t-t0)/2}}} = \frac{\Delta m_i}{\Delta t} \cdot \frac{1}{\gamma_{X_{V,t0}} + \frac{(\gamma_{X_{V,t}} - \gamma_{X_{V,t0}})}{2}} = \frac{\Delta m_i}{\frac{1}{2}(\gamma_{X_{V,t}} + \gamma_{X_{V,t0}}) \cdot \Delta t} \quad (18)$$

## Symbols Used

| | |
|---|---|
| $\boldsymbol{F}$ | Dynamic matrix |
| $\boldsymbol{H}$ | Observation matrix |
| $IVC$ | Integral of viable cells |
| $\boldsymbol{K}$ | Gain matrix |
| $\boldsymbol{l}$ | Vector of observations |
| $\Delta m_i$ | Change of mass of species $i$ |
| $\boldsymbol{n}$ | Dynamic noise of state vector |
| $\boldsymbol{P}^-$ | A priori covariance matrix |
| $\boldsymbol{P}^+$ | A posteriori covariance matrix |
| $\boldsymbol{Q}$ | Covariance matrix of the dynamic noise |

| | |
|---|---|
| $q_{i,I}$ | Specific rate of species $i$ calculated using the integral viable cell density |
| $q_{i,FD}$ | Specific rate of species $i$ calculated using the finite difference approximation |
| $\boldsymbol{R}$ | Covariance of observation noise |
| $U_r$ | Absolute uncertainty boundary of rate |
| $\Delta r_{rel}$ | Relative error of rate |
| $\Delta t$ | Time span used for deriving the rate |
| $\boldsymbol{w}$ | Observation noise |
| $\gamma_{X_V}$ | Viable cell concentration |
| $\boldsymbol{x}$ | State vector |
| $\boldsymbol{x}^-$ | A priori state vector |
| $\Delta y_i$ | Absolute measurement error of data point $i$ of the originating signal for rate calculation |

# References

[1]  P. Wechselberger, P. Sagmeister, H. Engelking, T. Schmidt, J. Wenger, C. Herwig, Bioprocess Biosyst. Eng. **2012**, 35 (9), 1637–1649. DOI:10.1007/s00449-012-0754-9.

[2]  P. Sagmeister, P. Wechselberger, C. Herwig, PDA J. Pharm. Sci. Technol. **2012**, 66 (6), 526–541. DOI:10.5731/pdajpst.2012.00889.

[3]  H.H. Ku, J. Res. Natl. Bur. Stand. Sect. C Eng. Instrum. **1966**, 70C (4), 263. DOI:10.6028/jres.070C.025.

[4]  P.R. Patnaik, AIChE J. **2004**, 50 (7), 1640–1646. DOI:10.1002/aic.10156.

[5]  J.E. Claes, J.F. Van Impe, Bioprocess Eng. **1999**, 21 (5), 389–395.

[6]  P.W. Sauer, J.E. Burky, M.C. Wesson, H.D. Sternard, L. Qu, Biotechnol. Bioeng. **2000**, 67 (5), 585–597.

[7]  M. Aehle, K. Bork, S. Schaepe, A. Kuprijanov, R. Horstkorte, R. Simutis, et al., Cytotechnology. **2012**, 64 (6), 623–634. DOI:10.1007/s10616-012-9438-1.

# Main mechanistic assumptions and equations

The main mechanistic assumptions behind data generation and soft-sensor are the same. Substrate, ammonia and oxygen is converted to biomass and carbon dioxide. As the amounts of products in biopharmaceutical processes are in the ranges of some milligrams per liter (CITE), the formed product can be neglected.

$$r_S CH_{pH}O_{pO} + r_{O2}O_2 + r_N NH_3 \rightarrow r_X CH_{zH}O_{zO}N_{zn} + r_{CO2}CO_2$$
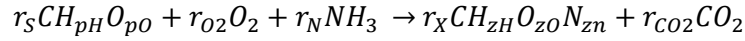
Main input signal into the model are the stoichiometry of the substrate ($C_6H_{12}O_6$) and the concentration (0.400 g mL$^{-1}$) and feed rate of the substrate. The stoichiometry of the organism was taken from CITE.

The feed rate at time point t during the not induced fed-batch phase was calculated in form as follows. After the induction phase, the feed rate was kept constant (Relevance shown in **Fehler! Verweisquelle konnte nicht gefunden werden.**).

$$F(t) = F_0 * e^{\mu t}$$

$t$ is the time (h), μ the feed exponent (h$^{-1}$) and $F_0$ the feed rate (mL) at fed-batch start which was calculated

$$F_0 = \frac{X * M_S * \mu}{Y_{X/S} * M_X * c_S}$$

Where $X$ is the total amount of biomass in the reactor (g), $M_X$ the C-normalized molecular weight of the biomass (g c-mol$^{-1}$), $Y_{X/S}$ the biomass substrate yield (c-mol c-mol$^{-1}$), $c_S$ the feed concentration (g mL$^{-1}$) and $M_S$ the C-normalized molecular weight of the substrate (g c-mol$^{-1}$).

As there is no substrate accumulation and no outflow of substrate during the fed-batch phase, the complete inflow of substrate is immediately consumed, resulting in a substrate uptake rate $r_S$ (c-mol h$^{-1}$) which is only dependent on the feed rate and the concentration of the feed.

$$r_S = -\frac{(F * c_S)}{M_S}$$

It should be mentioned here, that all rates where the flow direction shoes into the cell or where a species is consumed, were defined to be negative ($r_S$, OUR), while rates leading to an accumulation or formation of a species were defined to be positive ($r_X$, CER).

The biomass formation rate $r_X$ (c-mol h$^{-1}$) was calculated by using a fixed biomass/substrate yield in the exponential fed-batch phase, and a decreasing biomass/substrate yield in the induction phase.

$$r_X = -r_S * Y_{X/S}$$

The consumption of oxygen per formed amount of biomass $Y_{O2/X}$ (mol c-mol$^{-1}$) was calculated by setting up the electron balance. $\gamma_S$, $\gamma_S$ and $\gamma_{O2}$ are the degrees of reduction based on one c-mole of substrate and biomass (c-mol$^{-1}$), or one mole of oxygen (mol$^{-1}$), respectively. The degrees of reduction were calculated by setting up $\gamma_N = -3$, $\gamma_C = 4$, $\gamma_H = 1$ and $\gamma_O = -2$ (Villadsen, 2011). As $\gamma$ for $CO_2$, $NH_3$ and $H_2O$ according to the previous definition is 0, the degree of reduction in the system sum should not change over time.

106

$$r_S\gamma_S + r_X\gamma_X + r_{O_2}\gamma_{O_2} = 0$$

When setting $r_X$ to 1 and and $r_S$ to $1/Y_{X/S}$, $r_{O2}$ corresponds to $Y_{O2/X}$ and can be calculated according to the following equation:

$$Y_{O_2/X} = \frac{\dfrac{\gamma_S}{Y_{X/S}} - \gamma_X}{\gamma_{O_2}}$$

Using $Y_{O2/X}$, the oxygen uptake rate $OUR$ (mol h⁻¹) now can be simply calculated.

$$OUR = Y_{O_2/X} * r_X$$

For the calculation of the carbon dioxide evolution rate $CER$ (mol h⁻¹) it was assumed that the whole carbon flux goes into the biomass or leaves the reactor as carbon dioxide. When neglecting product formation and extracellular metabolites, the carbon balance can be stated as follows. All sum formulas are normalized to one carbon, resulting in the following equation:

$$r_S + r_X + r_{CO_2} = 0$$

As the accumulation of carbon dioxide in the reactor can be neglected, the carbon dioxide evolution rate was calculated as follows:

$$CER = r_{CO_2} = Y_{CO_2/S} * -r_S = (1 - Y_{X/S}) * -r_S$$

In the last step, the used oxygen and the produced carbon dioxide are added and subtracted from the inlet air and oxygen, considering water stripping and assuming the whole gas phase as ideal gas. The oxygen fraction in the air ($y_{O2, Air}$) is 0.2095, the oxygen fraction of the oxygen supply tank 0.9800 ($y_{O2, O2}$). The volumetric inflow of oxygen $O_{2, in}$ (L h⁻¹) is calculated as follows:

$$F_{O_2,in,total} = y_{O_2,Air} * F_{Air,in} + y_{O_2,O_2} * F_{O_2,in}$$

The volumetric oxygen outflow $F_{O2, out}$ (L h⁻¹) is

$$F_{O_2,out} = F_{O_2,in,total} + (OUR * V_M)$$

Similar for carbon dioxide $F_{CO2, out}$ (L h⁻¹)

$$F_{CO_2,out} = F_{CO_2,in} + (CER * V_M)$$

In the final step, the total outflow is calculated and the detected values $X_{CO2, out}$ and $X_{O2, out}$ (%) are generated. The value $y_{O2, wet}$ represents the oxygen content of the exhaust gas without microbial activity. It is an important value to estimate the water stripping effect and described in detail elsewhere (CITE). First the total outflow of air $F_{Air, out}$ has to be calculated

$$F_{Air,out} = \frac{F_{Air,in,total} + CER * V_M + OUR * V_M}{y_{O_2,wet}/y_{O_2,Air}}$$

$$X_{O_2,out} = 100 * \frac{F_{O_2,out}}{F_{Air,out}}$$

$$X_{CO_2,out} = 100 * \frac{F_{CO_2,out}}{F_{Air,out}}$$

## 9.3 Supporting Information: Criticality Assessment Workflow for Biopharmaceutical Process Validation Stage 1

*Table 1: Standardized experimental data from DoE study of primary recovery (PR), as well as upper and lower normal operating ranges (NOR_U, NOR_L, respectively) and scale down model (SDM) variance and mean. Normalization was performed by subtracting all values by the mean and diving by the standard deviation of DoE runs.*

| Batches | Parameter: temperature | Parameter: time | Parameter: Mixing [Yes/No] | Parameter: pH | Process impurity 2 concentration specific (post filtration) | Process impurity 1 concentration specific | Process impurity 2 concentration specific (prior filtration) |
|---|---|---|---|---|---|---|---|
| DoE1 | 0,00 | 1,22 | 0,95 | -1,22 | -0,88 | -1,20 | -0,55 |
| DoE2 | 1,22 | -1,22 | 0,95 | -1,22 | -0,81 | 0,80 | -0,63 |
| DoE3 | -1,22 | 0,00 | -0,95 | -1,22 | -0,74 | 0,06 | -0,23 |
| DoE4 | 0,00 | -1,22 | -0,95 | 1,22 | 0,87 | -0,88 | -0,39 |
| DoE5 | -1,22 | -1,22 | 0,95 | 0,00 | -0,71 | 0,93 | -0,41 |
| DoE6 | 1,22 | 1,22 | -0,95 | 0,00 | -0,76 | 0,51 | -0,66 |
| DoE7 | -1,22 | 1,22 | -0,95 | 1,22 | 2,07 | 0,68 | -0,24 |
| DoE8A | 0,00 | 0,00 | 0,95 | 0,00 | 0,17 | -1,25 | 1,98 |
| DoE9A | 0,00 | 0,00 | -0,95 | 0,00 | 0,92 | -1,03 | 1,76 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| DoE10 | 1,22 | 0,00 | 0,95 | 1,22 | -0,14 | 1,40 | -0,63 |
| Threshold | | | | | 78,76 | 24,41 | 77,31 |
| NOR_L | -1,71 | 0,33 | -0,95 | -0,61 | | | |
| NOR_U | 0,41 | 0,41 | 0,95 | 0,61 | | | |
| Sign Params | | | | | Parameter: pH | Parameter: pH, Parameter: temperature | |
| SDM_variance | | | | | 0,99 | 0,02 | 0,22 |
| SDM_mean | | | | | 3,27 | -0,70 | 0,98 |

*Table 2: Standardized experimental data from DoE study of chromatography column 1 (CC1), as well as upper and lower normal operating ranges (NOR_U, NOR_L, respectively) and scale down model (SDM) variance and mean. Normalization was performed by subtracting all values by the mean and diving by the standard deviation of DoE runs.*

| Batches | Parameter End pooling | Parameter Elution strength | Parameter wash strength | Parameter colum loading density | Parameter pH | Process impurity 2 clearance | Product impurity 2 clearance | Product impurity 1 clearance |
|---|---|---|---|---|---|---|---|---|
| DoE1 | 1.10 | -1.10 | -1.10 | 1.10 | -1.10 | -0.25 | -0.55 | -0.25 |
| DoE2 | 0.00 | 1.10 | -1.10 | -1.10 | -1.10 | -1.07 | -0.53 | -0.75 |
| DoE3 | -1.10 | 0.00 | -1.10 | 1.10 | 1.10 | -0.34 | -0.06 | -0.36 |
| DoE4 | -1.10 | -1.10 | -1.10 | -1.10 | 0.00 | -0.51 | 2.72 | -0.29 |
| DoE5 | 1.10 | -1.10 | 0.00 | -1.10 | 1.10 | 0.33 | NaN | NaN |
| DoE6 | 1.10 | 1.10 | -1.10 | 0.00 | 1.10 | -0.86 | -0.28 | -0.81 |
| DoE7 | -1.10 | 1.10 | 1.10 | -1.10 | 1.10 | -0.68 | 0.11 | -0.62 |
| DoE8 | -1.10 | 1.10 | 0.00 | 1.10 | -1.10 | -0.65 | -0.62 | 1.77 |
| DoE9 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.86 | -0.38 | -0.78 |
| DoE10 | 0.00 | -1.10 | 1.10 | 1.10 | 1.10 | 0.75 | 1.21 | 0.44 |
| DoE11 | -1.10 | -1.10 | 1.10 | 0.00 | -1.10 | 0.79 | -0.71 | 2.20 |
| DoE12 | 1.10 | 0.00 | 1.10 | -1.10 | -1.10 | -0.83 | -0.48 | 0.06 |
| DoE13 | 1.10 | 1.10 | 1.10 | 1.10 | 0.00 | 2.46 | -0.42 | -0.61 |
| Threshold | | | | | | -1.67 | -1.04 | -1.49 |
| NOR_U | 0.00 | 0.65 | 1.10 | 1.10 | 0.55 | | | |
| NOR_L | -1.10 | -1.10 | -1.10 | -0.51 | -0.55 | | | |

| Significant Parameters | | | | | | Parameter wash strength, Parameter column loading density | | Parameter end pooling, Parameter wash strength, Parameter column loading density, Parameter pH |
|---|---|---|---|---|---|---|---|---|
| SDM variance | | | | | | 0.03 | 0.00 | 0.02 |
| SDM mean | | | | | | -0.45 | -0.58 | -0.80 |

## 9.4 Supporting Information: Integrated Process Modeling – A process validation life cycle companion

*Table S3: Overview of found models based on DoE data. CC is abbreviation for chromatography column, PCI stands for process related impurities and PRI product related impurities.*

| | Response | alpha | Model | R² | Q² | ΔRQ | P | Parameters |
|---|---|---|---|---|---|---|---|---|
| **CC 1** | Yield | 0.05 | Linear | 0.88 | 0.74 | 0.14 | 0.000 | (+) End pooling, (+) Elution strength, (-) pH |
| | Specific PRI 1 clearance | 0.05 | Linear | 0.26 | 0.05 | 0.20 | 0.092 | (+) pH, |
| | Specific PRI 2 clearance | 0.05 | Linear | 0.77 | 0.50 | 0.27 | 0.006 | (-) pH, (+), Wash strength, (+) column loading density |
| | Specific DNA clearance | - | - | - | - | - | - | - |
| | Specific PCI 1 clearance | 0.05 | Quadratic | 0.91 | 0.62 | 0.30 | 0.003 | (+) End pooling, (-) End pooling^2, (+) Column loading density, (+) Wash strength, (-) pH |
| | Specific PCI 2 clearance | - | - | - | - | - | - | - |
| **CC 2** | Yield | - | - | - | - | - | - | - |
| | Specific PRI 1 clearance | - | - | - | - | - | - | - |
| | Specific PRI 2 clearance | - | - | - | - | - | - | - |
| | Specific DNA clearance | - | - | - | - | - | - | - |
| | Specific PCI 1 clearance | - | - | - | - | - | - | - |
| | Specific PCI 2 clearance | - | - | - | - | - | - | - |
| **CC 3** | Yield | 0.05 | Quadratic | 0.91 | 0.80 | 0.11 | 0.000 | (-) Column loading density^2, (+) Column loading density |
| | Specific PRI 1 clearance | 0.05 | Quadratic | 1.00 | 0.91 | 0.09 | 0.009 | (-) pH^2, (+) pH, (-) Wash strength, (+) Wash strength^2, (-) Column loading density^2, (+) Column loading density |
| | Specific PRI 2 clearance | - | - | - | - | - | - | - |
| | Specific DNA clearance | - | - | - | - | - | - | - |
| | Specific PCI 1 clearance | 0.05 | Quadratic | 0.99 | 0.98 | 0.02 | 0.000 | (+) Column loading density, (-) Column loading density^2, (+) Gradient slope |
| | Specific PCI 2 clearance | - | - | - | - | - | - | - |

*Table S4: Overview of models showing a correlation between specific CQA clearances and CQA load density. CC is abbreviation for chromatography column, PCI stands for process related impurities and PRI product related impurities.*

| | Response | alpha | model | R² | Q² | ΔRQ | P | Parameters |
|---|---|---|---|---|---|---|---|---|
| CC 1 | Specific PRI 1 clearance | - | - | - | - | - | - | - |
| | Specific PRI 2 clearance | - | - | - | - | - | - | - |
| | Specific DNA clearance | - | - | - | - | - | - | - |
| | Specific PCI 1 clearance | - | - | - | - | - | - | - |
| | Specific PCI 2 clearance | 0.05 | Linear | 0.78 | 0.68 | 0.10 | 0.000 | Load PCI 2 amount per CV |
| CC 2 | Specific PRI 1 clearance[1] | 0.05 | Linear | 0.66 | 0.42 | 0.24 | 0.000 | Load PRI 1 amount per CV |
| | Specific PRI 2 clearance | - | - | - | - | - | - | - |
| | Specific DNA clearance | - | - | - | - | - | - | - |
| | Specific PCI 1 clearance[2] | 0.05 | Linear | 0.54 | 0.36 | 0.18 | 0.000 | Load PCI 1 amount per CV |
| | Specific PCI 2 clearance | - | - | - | - | - | - | - |
| CC 3 | Specific PRI 1 clearance | - | - | - | - | - | - | - |
| | Specific PRI 2 clearance | - | - | - | - | - | - | - |
| | Specific DNA clearance | - | - | - | - | - | - | - |
| | Specific PCI 1 clearance | - | - | - | - | - | - | - |
| | Specific PCI 2 clearance | 0.05 | Linear | 0.63 | 0.37 | 0.26 | 0.002 | Load PCI 2 amount per CV |

[1] PRI 1 spiking experiments were used to establish this model
[2] PCI 1 depletion experiments were used to establish this model

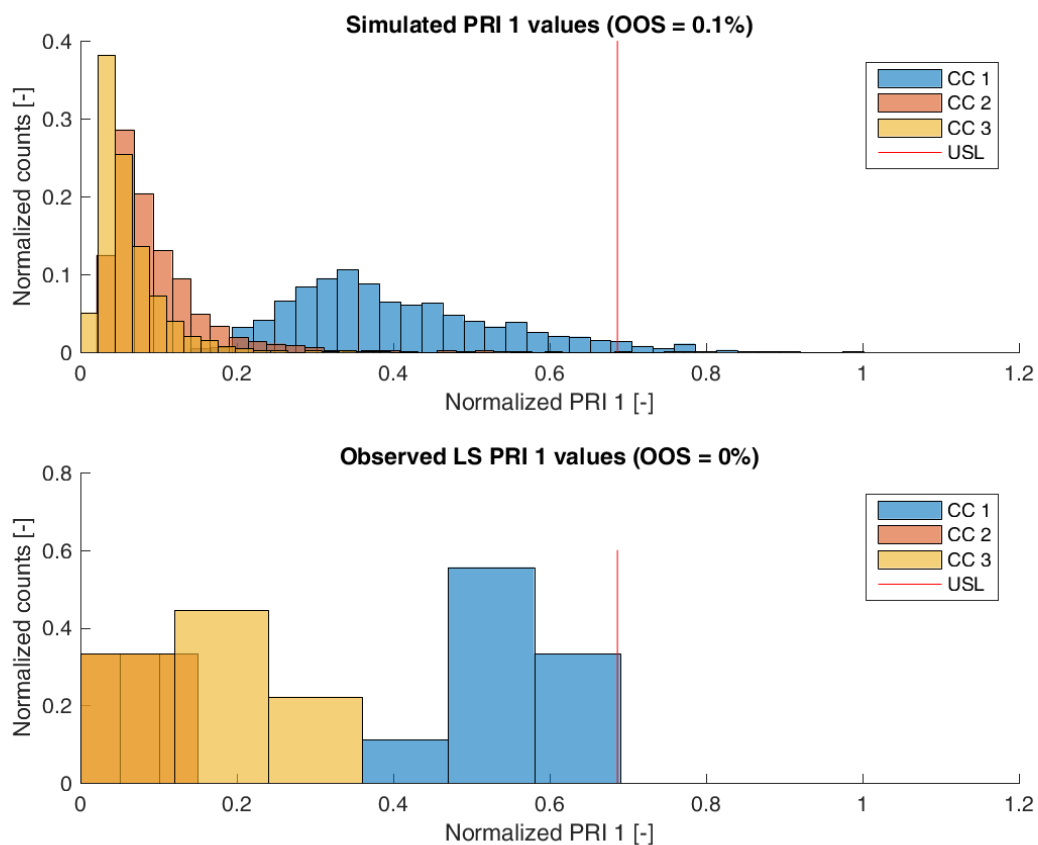# IPM Simulation without Spiking Models



*Figure S4: Comparison of simulated (top) product related impurity 1 distribution and observed (bottom) product related impurity 1 from LS after each column step. Normalization was performed by diving by the maximum observed $c_{CQA}$. Simulation was performed without taking any spiking model into account.*
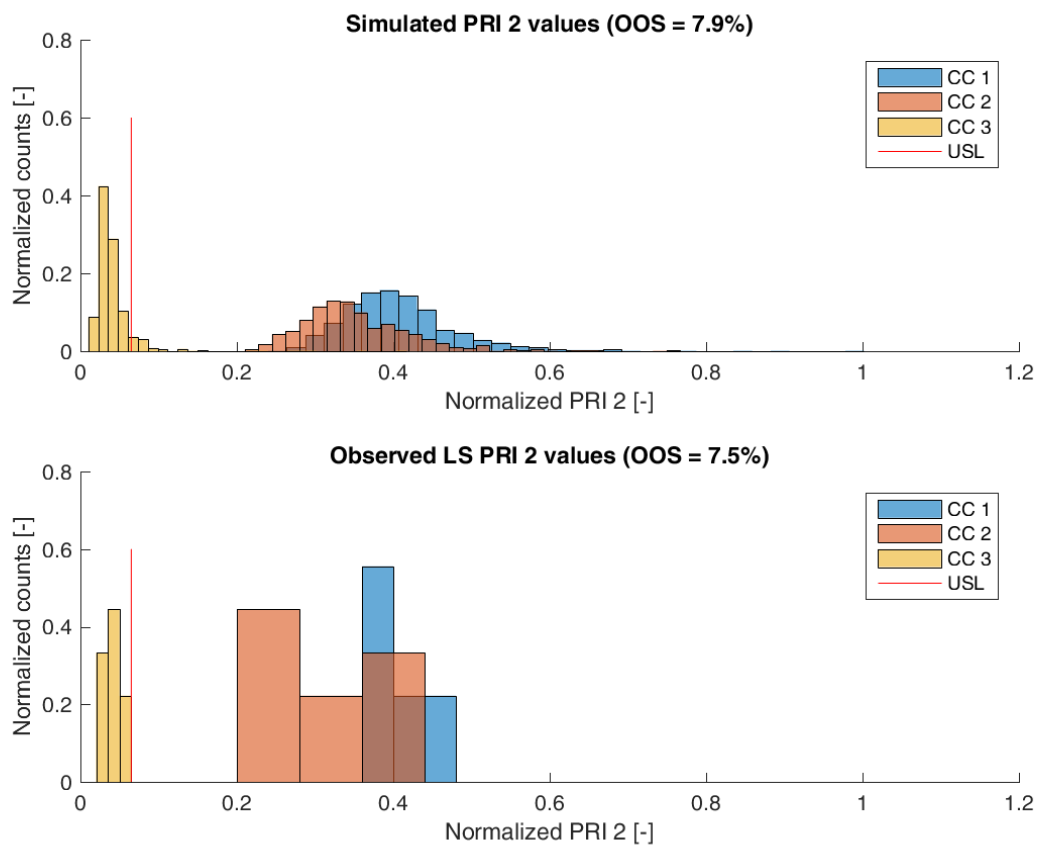
*Figure S5: Comparison of simulated (top) product related impurity 2 distribution and observed (bottom) product related impurity 2 from LS after each column step. Normalization was performed by diving by the maximum observed $c_{CQA}$. Simulation was performed without taking any spiking model into account.*
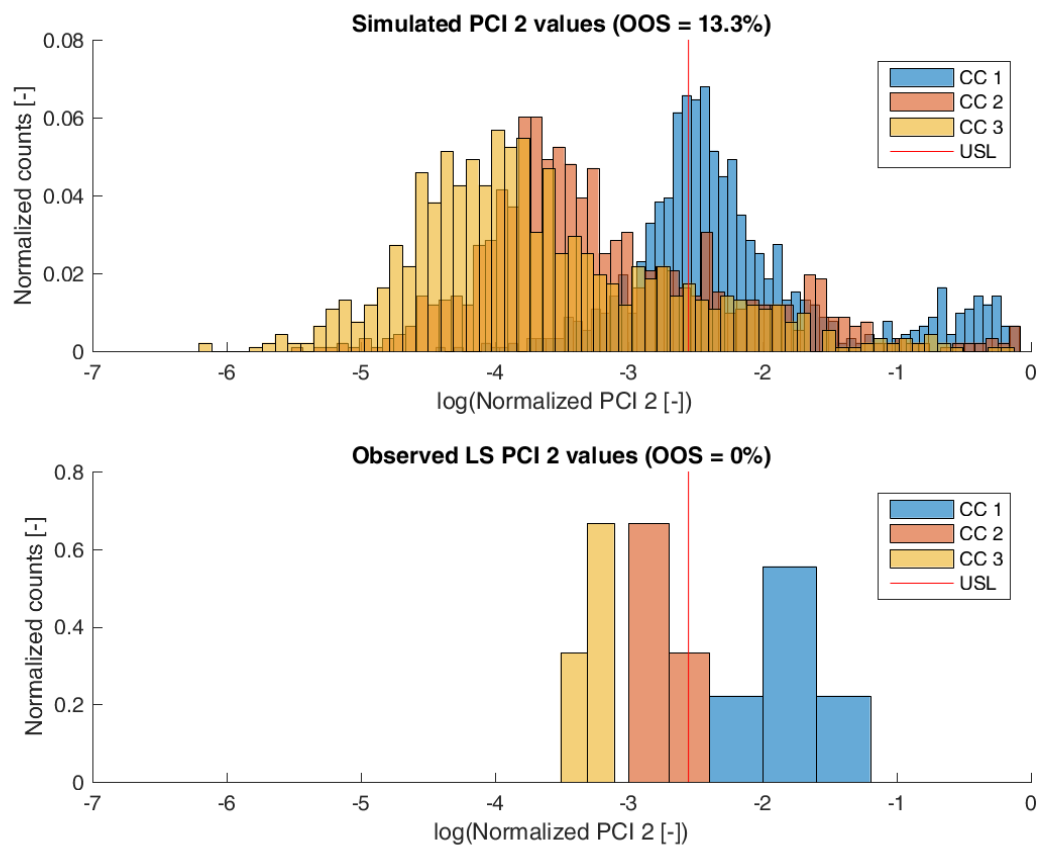
*Figure S6: Comparison of simulated (top) process related impurity 2 distribution and observed (bottom) process related impurity 2 from LS after each column step. Normalization was performed by diving by the maximum observed $c_{CQA}$. Simulation was performed without taking any spiking model into account.*
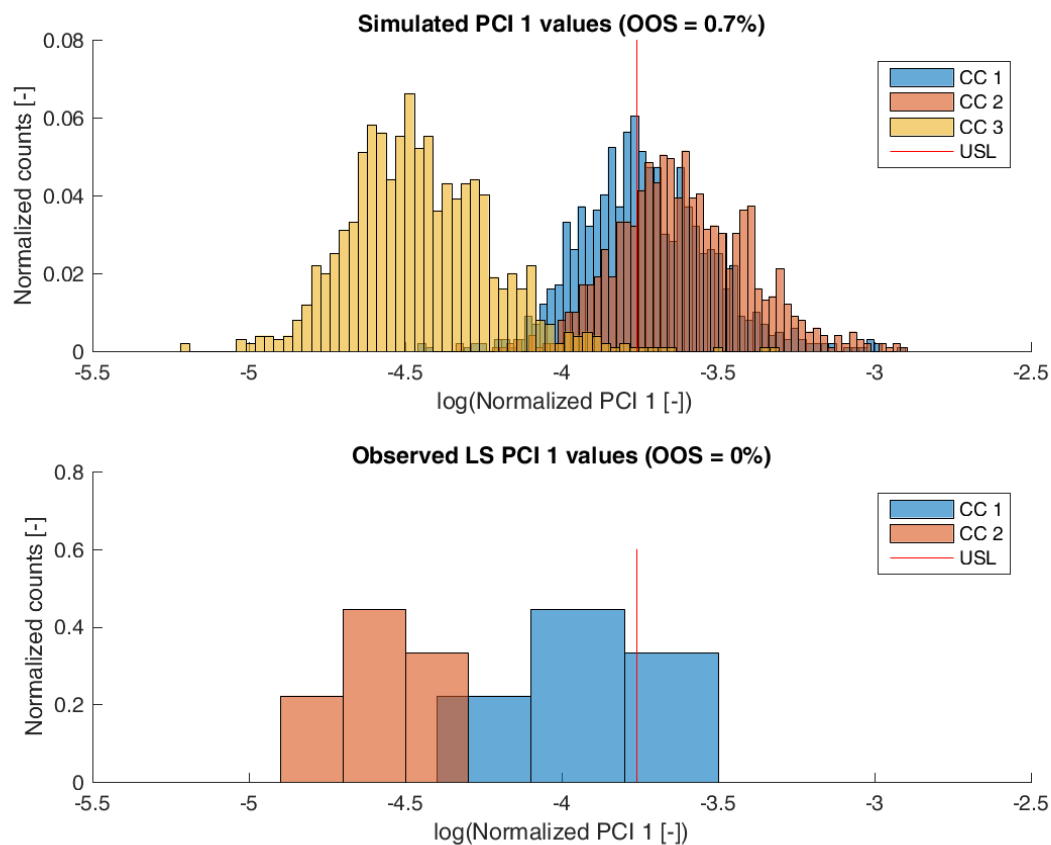
*Figure S7: Comparison of simulated (top) process related impurity 1 distribution and observed (bottom) process related impurity 1 from LS after each column step. For CC 3 pool, no process related impurity 1 value was observed above LoQ, therefore, no histogram bar is plotted for the observed values at CC 3 pool. Normalization was performed by diving by the maximum observed $c_{CQA}$. Simulation was performed without taking any spiking model into account.*